Information Systems 47 (2015) 98–115

Contents lists available at ScienceDirect

Information Systems

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

# The rise of "big data" on cloud computing: Review and open research issues

Ibrahim Abaker Targio Hashem [a,*], Ibrar Yaqoob [a], Nor Badrul Anuar [a], Salimah Mokhtar [a], Abdullah Gani [a], Samee Ullah Khan [b]

[a] Faculty of Computer Science and information Technology, University of Malaya, 50603 Kuala Lumpur, Malaysia
[b] NDSU-CIIT Green Computing and Communications Laboratory, North Dakota State University, Fargo, ND 58108, USA

ARTICLE INFO

*Article history:*
Received 11 June 2014
Received in revised form
22 July 2014
Accepted 24 July 2014
Recommended by: Prof. D. Shasha
Available online 10 August 2014

*Keywords:*
Big data
Cloud computing
Hadoop

ABSTRACT

Cloud computing is a powerful technology to perform massive-scale and complex computing. It eliminates the need to maintain expensive computing hardware, dedicated space, and software. Massive growth in the scale of data or big data generated through cloud computing has been observed. Addressing big data is a challenging and time-demanding task that requires a large computational infrastructure to ensure successful data processing and analysis. The rise of big data in cloud computing is reviewed in this study. The definition, characteristics, and classification of big data along with some discussions on cloud computing are introduced. The relationship between big data and cloud computing, big data storage systems, and Hadoop technology are also discussed. Furthermore, research challenges are investigated, with focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Lastly, open research issues that require substantial research efforts are summarized.



**Contents**



* Corresponding author. Tel.: +60 173946811.
*E-mail addresses:* targio@siswa.um.edu.my (I.A.T. Hashem), ibraryaqoob@siswa.um.edu.my (I. Yaqoob), badrul@um.edu.my (N.B. Anuar), salimah@um.edu.my (S. Mokhtar), abdullah@um.edu.my (A. Gani), samee.khan@ndsu.edu (S. Ullah Khan).

http://dx.doi.org/10.1016/j.is.2014.07.006
0306-4379/

## 1. Introduction

The continuous increase in the volume and detail of data captured by organizations, such as the rise of social media, Internet of Things (IoT), and multimedia, has produced an overwhelming flow of data in either structured or unstructured format. Data creation is occurring at a record rate [1], referred to herein as big data, and has emerged as a widely recognized trend. Big data is eliciting attention from the academia, government, and industry. Big data are characterized by three aspects: (a) data are numerous, (b) data cannot be categorized into regular relational databases, and (c) data are generated, captured, and processed rapidly. Moreover, big data is transforming healthcare, science, engineering, finance, business, and eventually, the society. The advancements in data storage and mining technologies allow for the preservation of increasing amounts of data described by a change in the nature of data held by organizations [2]. The rate at which new data are being generated is staggering [3]. A major challenge for researchers and practitioners is that this growth rate exceeds their ability to design appropriate cloud computing platforms for data analysis and update intensive workloads.

Cloud computing is one of the most significant shifts in modern ICT and service for enterprise applications and has become a powerful architecture to perform large-scale and complex computing. The advantages of cloud computing include virtualized resources, parallel processing, security, and data service integration with scalable data storage. Cloud computing can not only minimize the cost and restriction for automation and computerization by individuals and enterprises but can also provide reduced infrastructure maintenance cost, efficient management, and user access [4]. As a result of the said advantages, a number of applications that leverage various cloud platforms have been developed and resulted in a tremendous increase in the scale of data generated and consumed by such applications. Some of the first adopters of big data in cloud computing are users that deployed Hadoop clusters in highly scalable and elastic computing environments provided by vendors, such as IBM, Microsoft Azure, and Amazon AWS [5]. Virtualization is one of the base technologies applicable to the implementation of cloud computing. The basis for many platform attributes required to access, store, analyze, and manage distributed computing components in a big data environment is achieved through virtualization.

Virtualization is a process of resource sharing and isolation of underlying hardware to increase computer resource utilization, efficiency, and scalability.

The goal of this study is to implement a comprehensive investigation of the status of big data in cloud computing environments and provide the definition, characteristics, and classification of big data along with some discussions on cloud computing. The relationship between big data and cloud computing, big data storage systems, and Hadoop technology are discussed. Furthermore, research challenges are discussed, with focus on scalability, availability, data integrity, data transformation, data quality, data heterogeneity, privacy, legal and regulatory issues, and governance. Several open research issues that require substantial research efforts are likewise summarized.

The rest of this paper is organized as follows. Section 2 presents the definition, characteristics, and classification of big data. Section 3 provides an overview of cloud computing. The relationship between cloud computing and big data is presented in Section 4. Section 5 presents the storage systems of big data. Section 6 presents the Hadoop background and MapReduce. Several issues, research challenges, and studies that have been conducted in the domain of big data are reviewed in Section 7. Section 8 provides a summary of current open research issues and presents the conclusions. Table 1 shows the list of abbreviations used in the paper.

## 2. Definition and characteristics of big data

Big data is a term utilized to refer to the increase in the volume of data that are difficult to store, process, and analyze