Journal of Network and Computer Applications 71 (2016) 151–166

Contents lists available at ScienceDirect

Journal of Network and Computer Applications

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

A survey of big data management: Taxonomy and state-of-the-art

Aisha Siddiqa [a,b], Ibrahim Abaker Targio Hashem [a,b], Ibrar Yaqoob [a,b], Mohsen Marjani [a,b], Shahabuddin Shamshirband [b,*], Abdullah Gani [a,b], Fariza Nasaruddin [c]

[a] Center for Mobile Cloud Computing Research, University of Malaya, Kuala Lumpur, Malaysia
[b] Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia
[c] Department of Information System, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, Malaysia

ARTICLE INFO

Article history:
Received 22 November 2015
Received in revised form
12 February 2016
Accepted 11 April 2016
Available online 22 April 2016

Keywords:
Big data management
Storage
Big data
Processing
Security

ABSTRACT

The rapid growth of emerging applications and the evolution of cloud computing technologies have significantly enhanced the capability to generate vast amounts of data. Thus, it has become a great challenge in this big data era to manage such voluminous amount of data. The recent advancements in big data techniques and technologies have enabled many enterprises to handle big data efficiently. However, these advances in techniques and technologies have not yet been studied in detail and a comprehensive survey of this domain is still lacking. With focus on big data management, this survey aims to investigate feasible techniques of managing big data by emphasizing on storage, pre-processing, processing and security. Moreover, the critical aspects of these techniques are analyzed by devising a taxonomy in order to identify the problems and proposals made to alleviate these problems. Furthermore, big data management techniques are also summarized. Finally, several future research directions are presented.

1. Introduction

Over the last few years, the volume of data worldwide has exploded with the amplified use of various digital devices that continuously generate massive amounts of heterogeneous, structured or unstructured data, resulting in what is now called "big data" (Kambatla et al., 2014). Big data refers to rapidly growing amounts of data for which traditional database mechanisms have become inefficient in terms of storage, processing and analysis (Manyika et al., 2011). Managing big data with diverse data formats is a main basis for competition in business and management. Nonetheless, it has also become a new challenge for Information and Communication Technologies in both Science and industry encouraging the pursue of data-centric architectures and operational models (Han et al., 2014).

Meanwhile, traditional data storage and processing typically fed with relatively clean data sets generated by limited sources; hence, the results tended to be accurate. However, the evolution of big data has revealed a serious management problem, as standard tools and procedures are not designed to manage such massive data volumes (Philip Chen and Zhang, 2014). At the same time, current infrastructures are not yet capable of addressing the distributed computational needs of managing big data and exploiting large quantities and varieties of data (Candela et al., 2012). This is not only due to the growth in the volumes of data sets but also to their complexity and volatility that makes processing and analysis very hard to achieve through traditional data management techniques and technologies. Obviously, it is very challenging for current infrastructures to sustain huge amounts of data (Russom, 2011).

Current techniques and technologies designed to handle big data management problems mostly emphasize on certain characteristics of big data, such as volume, variety and velocity (Philip Chen and Zhang, 2014). Moreover, big data comprise complex data that are massively produced and managed in geographically dispersed repositories (Kambatla et al., 2014). Such complexity motivates the development of advanced management techniques and technologies for dealing with the challenges of big data. However, these advances in techniques and technologies have not yet been studied in detail and a comprehensive survey of this domain is required. Although several studies exist related to big data management (Han et al., 2014; Russom, 2013; McAfee and Brynjolfsson, 2012; Chaudhuri, 2012; Borkar et al., 2012), no one directly focused on technical aspects of big data management providing a description of existing techniques in storage, preprocessing, processing and security. Moreover, this survey analyzes several

* Corresponding author.
E-mail addresses: shamshirband1396@gmail.com (S. Shamshirband), abdullah@um.edu.my (A. Gani).

http://dx.doi.org/10.1016/j.jnca.2016.04.008
1084-8045/

problems inhibiting the big data management and review corresponding solutions by devising taxonomy.

This survey focuses primarily on big data aspects in the context of data management. A broad coverage of existing work on storage, preprocessing, processing and security is provided. In addition, this survey offers added value by means of a comprehensive taxonomy of existing techniques and technologies as well as highlighting the importance of typical big data management challenges related to storage, preprocessing, processing and security. Moreover, this survey aims to be a useful guide to challenges and solution in big data management and also a point of reference for future work on big data management. Furthermore, this survey summarizes the benefits that can be achieved if techniques are adopted for specific application areas of management, such as storage, preprocessing, analysis and/or security. Most significantly, this research contributes as a guide for researchers in the expedition of suitable big data management techniques and in the development of augmented techniques in response to the insufficiency of existing solutions.

In order to achieve the aims as mentioned above, we carry out our research investigation by answering the questions related to recent big data management advances as follows: (a) how big data management techniques optimize storage resources to meet rapid growth and fast retrieval requirements? (b) how pre-processing tools and technologies such as cleansing and transformation are managed to support upcoming trends of big data? (c) how big data analytics is being performed to deal with abundant information that could impact the business? (d) how security infers big data management process?

The contributions are as follows:

- A comprehensive review of big data management techniques with respect to data storage, pre-processing, processing and security
- A discussion on a taxonomy of big data management process flow with focus on the problems and available solutions related to storage, pre-processing, processing and security
- A comparison of different big data management techniques for storage, pre-processing and processing based on parameters including availability, scalability, integrity, heterogeneity, resource optimization and velocity
- A discussion on future directions and challenges regarding big data management

The rest of the article is organized as follows: Section 2 provides a general overview of big data management. Taxonomy of techniques for storage, pre-processing, processing and security aspects of big data is presented in Section 3. Section 4 discusses techniques for storage, pre-processing and processing and an analysis of their capability to meet big data management requirements, such as availability, scalability, integrity, heterogeneity, resource optimization and velocity. Section 5 highlights challenges and future research directions for big data management and Section 6 concludes the study.

2. Overview of big data management

Big data management is a new discipline, where data management techniques, tools and platforms including storage, preprocessing, processing and security can be applied. However, data management is a broad practice that encompasses other data disciplines, such as data warehousing, data integration, data quality and data governance (McAfee and Brynjolfsson, 2012). Thus, big data management is a complex process, particularly when abundant data originating from heterogeneous sources are to be used for business intelligence and decision-making (Baker, 2014). Furthermore, big data management has become a key to the success of many enterprises, science, industries, engineering fields and government ventures (Chaudhuri et al., 2011). The main objective is to enhance data quality and accessibility for decision-making and improve productivity. Russom (2013) reported in a survey on managing big data that 75% of organizations manage some form of big data.

Big data management has been successfully adopted in several big data techniques and technologies to support its processes. Big data management comes with new challenges in terms of data integration complexity, storage capacity, analytical tools and lack of governance (Russom, 2013). Han and Yonggang (Han et al., 2014) categorized big data management processes into (i) big data science and (ii) big data infrastructure. Big data science refers to the study of techniques and technologies regarding big data acquisition, conditioning and evaluation. Techniques and technologies are developed with attention to improve existing methods of managing, analyzing, visualizing and exploiting informative knowledge from various ample data (Philip Chen and Zhang, 2014). On the other hand, big data infrastructure is derived from more than one big data framework to enable processing large amounts of data in distributed environments across clusters of machines (Han et al., 2014).

Fig. 1 depicts the scenario of healthcare system with the perspective of big data management. The aim of the system is to generate the report from the voluminous amount of data stored in the HDFS. Generally, when authorized medical staff submit queries, before getting the required information from pool of data, the data passes through multiple steps. First, preprocessing methods such as cleansing, integration and transmission of data to make them ready for processing are applied. After cleansing process, mining techniques are applied to extract the required valuable information in the form of reports. Encryption techniques are used during communication of the storage, pre-processing and processing modules.

The work most relevant to the current survey is undoubtedly recent work on the big data management revolution (McAfee and Brynjolfsson, 2012). However, this survey provides a more in-depth investigation of the challenges and solutions related to big data management besides devising taxonomy. Related works include: What next?: a half-dozen data management research goals for big data and the cloud (Chaudhuri, 2012) and Inside Big Data

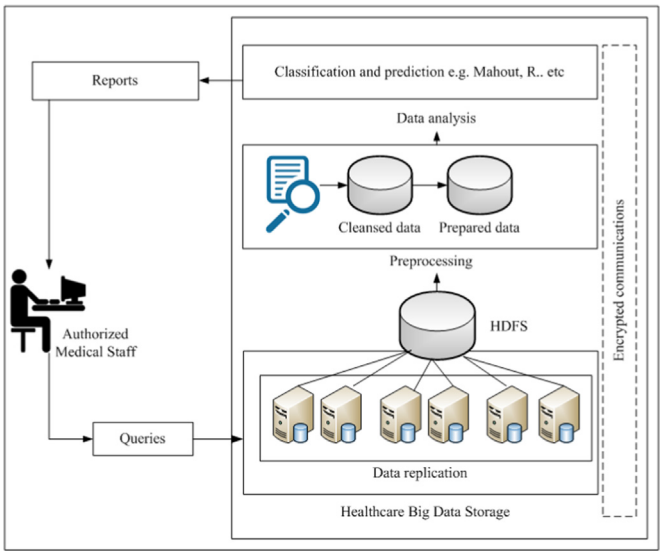

Fig. 1. Healthcare big data management scenario.