Computer Networks 132 (2018) 180–199

Contents lists available at ScienceDirect

Computer Networks

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

Review article

Big data analytics for wireless and wired network design: A survey

Mohammed S. Hadi*, Ahmed Q. Lawey, Taisir E.H. El-Gorashi, Jaafar M.H. Elmirghani

School of Electronic and Electrical Engineering, University of Leeds, United Kingdom

ARTICLE INFO

Article history:
Received 16 June 2017
Revised 13 January 2018
Accepted 16 January 2018

Keywords:
Big data analytics
Network design
Self-optimization
Self-configuration
Self-healing network

ABSTRACT

Currently, the world is witnessing a mounting avalanche of data due to the increasing number of mobile network subscribers, Internet websites, and online services. This trend is continuing to develop in a quick and diverse manner in the form of big data. Big data analytics can process large amounts of raw data and extract useful, smaller-sized information, which can be used by different parties to make reliable decisions.

In this paper, we conduct a survey on the role that big data analytics can play in the design of data communication networks. Integrating the latest advances that employ big data analytics with the networks' control/traffic layers might be the best way to build robust data communication networks with refined performance and intelligent features. First, the survey starts with the introduction of the big data basic concepts, framework, and characteristics. Second, we illustrate the main network design cycle employing big data analytics. This cycle represents the umbrella concept that unifies the surveyed topics. Third, there is a detailed review of the current academic and industrial efforts toward network design using big data analytics. Forth, we identify the challenges confronting the utilization of big data analytics in network design. Finally, we highlight several future research directions. To the best of our knowledge, this is the first survey that addresses the use of big data analytics techniques for the design of a broad range of networks.

1. Introduction

Networks generate traffic in rapid, large, and diverse ways, which leads to an estimate of 2.5 exabytes created per day [1]. There are many contributors to the increasing size of the data. For instance, scientific experiments can generate lots of data, such as CERN's Large Hadron Collider (LHC) that generates over 40 petabyte each year [2]. Social media also has its share, with over 1 billion users, spending an average 2.5 h daily, liking, tweeting, posting, and sharing their interests on Facebook and Twitter [3]. It is without a doubt that using this activity-generated data can affect many aspects, such as intelligence, e-commerce, biomedical, and data communication network design. However, harnessing the powers of this data is not an easy task. To accommodate the data explosion, data centers are being built with massive storage and processing capabilities, an example of which is the National Security Agency (NSA) Utah data centre that can store up to 1 yottabyte of data [4], and with a processing power that exceeds 100 petaflops [5]. Due to the increased needs to scale-up databases to data volumes that exceeded processing and/or storage capabilities, systems that ran on computer clusters started to emerge. Perhaps the first milestone took place in June 1986 when Teradata [6] used the first parallel database system (hardware and software), with one terabyte storage capacity, in Kmart data warehouse to have all their business data saved and available for relational queries and business analysis [7,8]. Other examples include the Gamma system of the University of Wisconsin [9] and the GRACE system of the University of Tokyo [10].

In light of the above, the term "Big Data" emerged, and it can be defined as high-volume, high-velocity, and high-variety data that provides substantial opportunities for cost-effective decision-making and enhanced insight through advanced processing which extracts information and knowledge from data [11]. Another way to define big data is by saying it is the amount of data that is beyond traditional technology capabilities to store, manage, and process in an efficient and easy way [12]. Big data is already being employed by digital-born companies like Google and Amazon to help these companies with data-driven decisions [13]. It also helps in the development of smart cities and campuses [14], as well as in other fields like agriculture, healthcare, finance [15], and transportation [16]. Big data has the following characteristics:

1- *Volume*: This is a representation of the data size [17].

* Corresponding author.
E-mail addresses: elmsha@leeds.ac.uk (M.S. Hadi), a.q.lawey@leeds.ac.uk (A.Q. Lawey), T.E.H.Elgorashi@leeds.ac.uk (T.E.H. El-Gorashi), J.M.H.Elmirghani@leeds.ac.uk (J.M.H. Elmirghani).

https://doi.org/10.1016/j.comnet.2018.01.016
1389-1286/

Table 1
Various big data dimensions.

No. of Vs	References	Dimensions (Characteristics)								
		Volume	Velocity	Variety	Veracity	Value	Variability	Volatility	Validity	Complexity
3Vs	[25–31]	✓	✓	✓						
4Vs	[4,32–34]	✓	✓	✓	✓					
	[35–39]	✓	✓	✓		✓				
5Vs	[3,11,21,40,41]	✓	✓	✓	✓	✓				
6Vs	[20,22,24,42]	✓	✓	✓	✓	✓	✓			✓
7Vs	[23,43]	✓	✓	✓	✓	✓		✓	✓	

2- *Variety*: Generating data from a variety of sources results in a range of data types. These data types can be structured (e.g. e-mails), semi-structured (e.g. log files data from a webpage); and unstructured (e.g. customer feedback), and hybrid data [18].
3- *Velocity*: Is an indication of the speed of the data when being generated, streamed, and aggregated [19]. It can also refer to the speed at which the data has to be analyzed to maintain relevance [17].

Depending on the research area and the problem space, other terms or Vs can be added. For example, is this data of any value? How long can we consider this an accurate and valid data? Since we are conducting a survey, we find it compelling to briefly introduce other Vs as well. Typically, the number of analyzed Vs is 3 to 7 in a single paper (e.g. 6V+C [20]), where C represents *Complexity*, however, different papers analyze different sets of Vs and the union (sum) of all the analyzed Vs among all surveyed papers is 8V and a C, as shown in Table 1.

4- *Value*: Is a measure of data usefulness when it comes to decision making [19], or how much added-value is brought by the collected data to the intended process, activity, or predictive analysis/hypothesis [21].
5- *Veracity*: Refers to the authenticity and trustworthiness of the collected data against unauthorized access and manipulation [21,22].
6- *Volatility*: An indication of the period in which the data can still be regarded as valid and for how long that data should be kept and stored [23].
7- *Validity*: This might appear similar to veracity; however, the difference is that validity deals with data accuracy and correctness regarding the intended usage. Thus, certain data might be valid for an application but invalid for another.
8- *Variability*: This refers to the inconsistency of the data. This is due to the high number of distributed autonomous data sources [24]. Other researchers refer to the variability as the consistency of the data over time [22].
9- *Complexity*: A measure of the degree of interdependence and inter-connectedness in big data [20]. Such that, a system may witness a (substantial, low, or no) effect due to a very small change(s) that ripples across the system [19]. Also, complexity can be considered in terms of relationship, correlation and connectivity of data. It can further manifest in terms of multiple data linkages, and hierarchies. Complexity and its mentioned attributes can however help better organize big data. It should be noted that complexity was included among the big data attributes (Vs) in [20] where big data was characterized as having 6V + complexity. This is how we will arrange it in Table 1.

The process of extracting hidden, valuable patterns, and useful information from big data is called *big data analytics* [44]. This is done through applying advanced analytics techniques on large data sets [28]. Before commencing the analytics process, data sets may comprise certain consistency and redundancy problems affecting their quality. These problems arise due to the diverse sources from which the data originated. *Data pre-processing* techniques are used to address these problems. The techniques include integration, cleansing (or cleaning), and redundancy elimination, and they were discussed by the authors in [39].

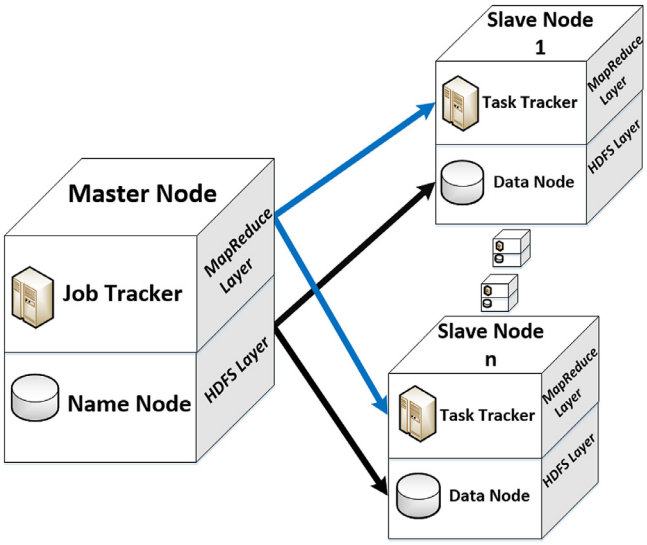

Fig. 1. Hadoop V1.x architecture.

Big data analytics can be carried out using a number of frameworks (shown below) that usually require an upgradeable cluster dedicated solely for that purpose [17]. Even if the cluster can be formed using a number of commodity servers [45], however, this still forms an impediment for limited-budget users who want to analyze their data. The solution is presented through the democratization of computing. This made it possible for any-sized company and business owners to analyze their data using cloud computing platforms for big data analytics. Consequently, the use of big data analytics is not limited to enterprise-level companies. Furthermore, business owners do not have to heavily invest in an expensive hardware dedicated to analyzing their data [1]. Amazon is one of the companies that provide 'cloud-computed' big data analytics for its customers. The service is called Amazon EMR (Elastic MapReduce), and it enables users to process their data in the cloud with a considerably lower cost in a pay-as-you-use fashion. The user is able to shrink or expand the size of the computing clusters to control the data volume handled and response time [1,46]

Dealing with big amounts of data is not an easy task, especially if there is a certain goal in mind since data arrives in a fast manner, it is vital to provide fast collection, sorting, and processing speeds. Apache Hadoop was created by Doug Cutting [47] for this purpose. It was later adopted, developed, and released by Yahoo [48]. Apache Hadoop can be defined as a top-level, java-written, open source framework. It utilizes clusters of commodity hardware [49].