# Chapter 102
# Intelligent Management and Efficient Operation of Big Data

**José Moura**
*Instituto Universitario de Lisboa, Portugal & Instituto de Telecomunicações, Portugal*

**Fernando Batista**
*Instituto Universitario de Lisboa, Portugal*

**Elsa Cardoso**
*Instituto Universitario de Lisboa, Portugal*

**Luís Nunes**
*Instituto Universitario de Lisboa, Portugal & Instituto de Telecomunicações, Portugal*

## ABSTRACT

*This chapter details how Big Data can be used and implemented in networking and computing infrastructures. Specifically, it addresses three main aspects: the timely extraction of relevant knowledge from heterogeneous, and very often unstructured large data sources; the enhancement on the performance of processing and networking (cloud) infrastructures that are the most important foundational pillars of Big Data applications or services; and novel ways to efficiently manage network infrastructures with high-level composed policies for supporting the transmission of large amounts of data with distinct requisites (video vs. non-video). A case study involving an intelligent management solution to route data traffic with diverse requirements in a wide area Internet Exchange Point is presented, discussed in the context of Big Data, and evaluated.*

## 1. INTRODUCTION

Big Data is a relatively new concept. When someone is asked to define it, the tale of the blind man and the elephant immediately comes to mind. As in the tale, each person that talks about Big Data seems to have his/her own view, according to the person's background or the intended use of the data (Ward & Barker, 2013; McAfee & Brynjolfsson, 2012; Cox & Ellsworth, 1997; Diebold, 2012; Press, 2013). Big Data is closely related to the area of analytics (Davenport et al., 2010), as it also seeks to gather intel-

DOI: 10.4018/978-1-5225-7501-6.ch102

ligence from data generating value to the business or the organization. However, a Big Data application differs in terms of the volume (referring to large data volumes), velocity (i.e., multi-structured data types) and variety (related to the change rate and time-sensitive usage to maximize the business value) of the data involved. These aspects are usually known as the 3V's. These large and diverse data streams require "ever-increasing processing speeds, yet must be stored economically and fed back into business-process life cycles in a timely manner," (Michael & Miller, 2013, pp. 22). Big Data applications offer new opportunities of information processing for enhanced insight and decision-making in different disciplines such as business, finance, healthcare, transportation, research, and politics.

The successful deployment of a Big Data infrastructure requires the extraction of relevant knowledge from original heterogeneous (Parise et al, 2012), highly complex (Nature, 2008) and massive amount of data. To this end, several tools from different areas can be applied: Business Intelligence (BI) and Online Analytical Processing (OLAP), Cluster Analysis, Crowdsourcing, Network Analysis, Text Mining, and Natural Language Processing (NLP). As an example, massive amounts of textual information are constantly being produced and can be accessed from online sources, including social networks, blogs, and numerous websites. Such unstructured texts represent potentially valuable knowledge for companies, organizations, and governments. The process of extracting useful information from such unstructured texts, known as Text Mining, is now becoming a relevant research area. It draws from different fields of computer science, such as Web Mining, Information Retrieval (IR), NLP, Machine Learning (ML), and Data Mining. Today's text mining research and technology enables high-performance analytics from web's textual data, allowing to: cluster documents and web pages according to their content, find associations among entities (people, places and/or organizations), and reasoning about important data trends.

The data sets in Big Data are becoming increasingly complex (Nature, 2008). For example, the biology field is urging for robust data computing (The Apache Software Foundation, 2014a) and distributed storage solutions (The Apache Software Foundation, 2014c); machine learning algorithms for data mining tasks (Hall et al., 2009); online community collaborations need wiki-style information cooperative tools (Waldrop, 2008); sophisticated visualization techniques of intracellular signaling pathways require tools like GenMAPP (Waldrop, 2008); and innovative ways to control the Big Data infrastructure such as software-design networking (SDN). To conclude, Lawrence Hunter, a biological researcher, wrote: "Getting the most from the data requires interpreting them in light of all the relevant prior knowledge," (Marx, 2013). Clearly, satisfying this requisite also demands for new scalable Big Data solutions. In this way, the Big Data is a very challenging and exciting research area to be further explored and investigated.

An important aspect to guarantee the success of Big Data solutions is to manage with more intelligence the supporting computing/networking infrastructure. Currently, both data and collaborative applications are increasingly being moved towards Data Centers aggregated inside the cloud. Consequently, to obtain a good performance in Big Data applications it is mandatory to achieve a proper performance in Data Centers. To achieve this, some management enhancements are needed to operate more intelligently the available resources, such as: virtual machines (Dai et al., 2013), memory (Zhou & Li, 2013), CPU scheduling (Bae et al., 2012), cache (Koller et al., 2011), I/O (Ram et al., 2013), and network (Marx, 2013; Lange et al., 2011; Saleem, Hassan, & Asirvadam, 2011).

This chapter details how Big Data can be deployed and used, focusing on the following important aspects: i) timely extraction of relevant knowledge from heterogeneous, and often unstructured data sources; i) enhancement of the performance of processing and networking (cloud) infrastructures for Big Data, using more intelligent management solutions/algorithms; and iii) SDN as an intelligent and