Chapter 6
The Big Data Era: Data Management Novelties for Visualizing, Exploring, and Processing Big Data

Maria K. Krommyda
National Technical University of Athens, Greece

Verena Kantere
National Technical University of Athens, Greece

ABSTRACT

Large datasets pertaining to many scientific fields and everyday activities are becoming available at an increasing rate. Processing, analyzing, and understanding the information that they offer poses significant technical challenges. There are many efforts dedicated to the development of big data exploration, analysis, and visualization applications that will improve the value of the information extracted from these datasets. An analysis of the state-of-the-art in these applications is presented here along with open research challenges that have not yet been tackled sufficiently. Also, specific domains where big data applications are needed are presented, and unique challenges are identified.

INTRODUCTION

This chapter presents the concept of Big Data, discusses their unique characteristics and the way their definition has evolved in time. Next, key domains where their analysis provides additional knowledge and supports decision making are presented. Then, the chapter focuses on two Big Data research fields that try to address the needs of a wider audience, people that can greatly benefit from the available datasets but without access to the needed infrastructure. This chapter presents Big Data exploration and visualization applications that necessitate the development of methods and techniques that can make datasets accessible, and the current state of the research along with open research challenges. The objective of this chapter is to provide the reader with a deep understanding of the Big Data era, emphasize their

DOI: 10.4018/978-1-7998-3499-1.ch006

unique characteristics and explain how these contribute to their importance and potential. In addition, the chapter offers to the reader a clear understanding of the current research state, the areas that can be further explored and the expected next steps for the research of the field.

There are many different definitions for the term Big Data (De Mauro, 2015; De Mauro, 2016; Ward, 2013), depending on the time they were written and the field that they are referring to, but they all agree that there are large, complex and unprocessed datasets, that cannot be processed by traditional application but can offer knowledge and value if properly analyzed. Initially, there was a controversy regarding the volume of the dataset and what should be considered large or difficult to process.

This was mainly due to the fact that research has shown (Hilbert, 2011) that the application-specific capacity of the machines to compute information per capita has roughly doubled every 14 months, whereas the world's storage capacity per capita required roughly 40 months to double during the last decades.

The exponential growth of the data production, the diversity of the data sources, along with the improvement of the computational capabilities of the hardware made the quantification of the term insignificant and added multiple dimensions to the problem. To this end, a dataset is now characterized as Big Data when it complies with the seven Vs rule (Ali-ud-din Khan, 2014; Hilbert, 2016).

- **Volume.** This rule begins with the amount of data as it is a very important aspect of Big Data, given that the goal is to process high volumes of low-density, unstructured data. However, high volume is not defined in an absolute and catholic manner, but it is rather specific to each consumer of information, and it is defined by the quantity of the existing or generated data, the storage capabilities and the user's ability to process and analyze data. For some applications, this might be as low as tens of terabytes of data while for others it may easily reach hundreds of petabytes.
- **Variety.** It refers to the type and nature of the data. Big Data distance themselves from structured data schemas that fit neatly in relational databases. Nowadays, the majority of the raw data are available in unstructured and semi-structured data types including video, audio or metadata, such as clickstreams on a webpage. These data introduce the overhead of additional pre-processing to derive the respective information.
- **Variability.** This characteristic is different from 'Variety', as it focuses on the intended purpose of data usage. It is a characteristic, mostly associated with Natural Language Processing applications, as it emphasizes on properly interpreting the meaning of raw data taking into account the overall context. A very indicative example are words the meaning of which can completely change depending on the context they are used. The word 'sanction' is such an example, as it can mean a threatened penalty for disobeying a law or official permission for an action, making its interpretation reliable only when based entirely on the context in which it is used.
- **Velocity.** In this context there are two different aspects of velocity. On one hand, with the term velocity we address the speed at which the data are generated and/or updated. This identifies the rate that the system should achieve for the data update and challenges of data storage in order to avoid any loss of information. Big Data are expected to be produced continuously. On the other hand, Big Data analysis is expected to achieve near real-time results, so velocity also refers to the frequency of handling, processing, and publishing the results of the data manipulation.
- **Veracity.** It is one of the characteristics that was added later in the definition of Big Data and refers to the data accuracy. The sheer volume of the data available combined with the diversity of the data sources and the lack of official requirements for control often produce datasets that contain data that are inaccurate or falsified.