Technological Forecasting & Social Change xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Technological Forecasting & Social Change

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

Big Data sources and methods for social and economic analyses

Desamparados Blazquez, Josep Domenech*

Department of Economics and Social Sciences, Universitat Politècnica de València, Camí de Vera s/n., Valencia 46022, Spain

ARTICLE INFO

Keywords:
Big Data architecture
Forecasting
Nowcasting
Data lifecycle
Socio-economic data
Non-traditional data sources
Non-traditional analysis methods

ABSTRACT

The Data Big Bang that the development of the ICTs has raised is providing us with a stream of fresh and digitized data related to how people, companies and other organizations interact. To turn these data into knowledge about the underlying behavior of the social and economic agents, organizations and researchers must deal with such amount of unstructured and heterogeneous data. Succeeding in this task requires to carefully plan and organize the whole process of data analysis taking into account the particularities of the social and economic analyses, which include the wide variety of heterogeneous sources of information and a strict governance policy. Grounded on the data lifecycle approach, this paper develops a Big Data architecture that properly integrates most of the non-traditional information sources and data analysis methods in order to provide a specifically designed system for forecasting social and economic behaviors, trends and changes.

1. Introduction

What comes to your mind when talking about “The Digital Era”? For sure, concepts as the “Internet”, “Smartphones” or “Smart sensors” arise. These technologies are progressively being used in most of the everyday activities of companies and individuals. For instance, many companies conduct marketing campaigns through social networks, sell their products online, monitor the routes followed by sales representatives with smartphones or register the performance of machinery with specific sensors. At the other side, individuals make use of computers, smartphones and tablets in order to buy products online, share their opinions, chat with friends or check the way to some place. Moreover, citizens' movements and activities are daily registered by sensors placed in any part of cities or roads and in public places such as supermarkets.

Therefore, all of these technologies are generating tons of digitized and fresh data about people and firms' activities that properly analyzed, could help reveal trends and monitor economic, industrial and social behaviors or magnitudes. These data are not only updated, but also massive, given that daily data generation has been recently estimated in 2.5 Exabytes (IBM, 2016). For this reason, they are commonly referred to as “Big Data”, concept which first appeared in the late 90s (Cox and Ellsworth, 1997) and was defined in the early 2000s in terms of the 3Vs model (Laney, 2001), which refers to: Volume (size of data), Velocity (speed of data transfers), and Variety (different types of data, ranging from video to data logs for instance, and with different structures). This model evolved to adapt to the changing digital reality, so that it was extended to 4Vs, adding the “Value” dimension (process to extract valuable information from data, known as Big Data Analytics). Currently, the “Big Data” concept is starting to be defined in terms of the 5Vs model (Bello-Orgaz et al., 2016), which added the “Veracity” dimension (related to proper data governance and privacy concerns).

This new data paradigm is called to transform the landscape for socio-economic policy and research (Einav and Levin, 2014; Varian, 2014) as well as for business management and decision-making. Thus, identifying which data sources are available, what type of data they provide, and how to treat these data is basic to generate as much value as possible for the company or organization. In this context, a Big Data architecture adapted to the specific domain and purpose of the organization contributes to systematize the process of generating value. This architecture should be capable of managing the complete data lifecycle in the organization, including data ingestion, analysis and storage, among others.

Furthermore, the design of a Big Data architecture should consider the numerous challenges that this paradigm implies. These include: scalability, data availability, data integrity, data transformation, data quality, data provenance (related to generation of right metadata that identify the origin of data as well as the processes applied to them during the data lifecycle, to assure traceability), management of huge volumes of information, data heterogeneity (structured and unstructured, with different time frequencies), integration of data from different sources, data matching, bias, availability of tools for properly analyzing such kind of data, processing complexity, privacy and legal issues, and data governance (Fan et al., 2014; Jagadish et al., 2014;

* Corresponding author.
E-mail addresses: mdeblzso@upvnet.upv.es (D. Blazquez), jdomenech@upvnet.upv.es (J. Domenech).

http://dx.doi.org/10.1016/j.techfore.2017.07.027
Received 1 March 2017; Received in revised form 7 July 2017; Accepted 25 July 2017

Hashem et al., 2015).

The Big Data paradigm also offers many advantages and benefits for the companies, governments, and the society. Jin et al. (2015) highlight its potential contribution to national and industrial development, as it enforces to change and upgrade research methods, promotes and makes it easy to conduct interdisciplinary research, helps to nowcast the present and to forecast the future more precisely. In this vein, first Big Data architectures designed for specific fields are being proposed in order to surpass the previously mentioned challenges and make the most of the data available with the aim of nowcasting and forecasting variables of interest.

However, no specific architecture for social and economic forecasting has been proposed yet. This emerges as a necessity, in the one hand, because of the particular nature of socio-economic data, which have important components of uncertainty and human behavior that are particularly complex to model; and, in the other hand, because of the great benefits that can be derived from the use of Big Data to forecast economic and social changes. For instance, Big Data approaches have been proved to improve predictions of economic indicators such as the unemployment level (Vicente et al., 2015), help managers detect market trends so that they can anticipate opportunities, and also help policy-makers monitor faster and more precisely the effects of a wide range of policies and public grants (Blazquez and Domenech, 2017).

In this context, this paper aims to i) establish a framework about the new and potentially useful available sources of socio-economic data and new methods devoted to deal with these data, ii) propose a new data lifecycle model that encompasses all the processes related to working with Big Data, and iii) propose an architecture for a Big Data system able to integrate, process and analyze data from different sources with the objective to forecast economic and social changes.

The remainder of the paper is organized as follows: Section 2 reviews the Big Data architectures proposed in the literature; Section 3 compiles the new socio-economic data sources emerged in the Digital Era and proposes a classification of them; Section 4 reviews the new methods and analytics designed to deal with Big Data and establishes a taxonomy of these methods; Section 5 depicts the data lifecycle on which the proposed Big Data architecture is based; Section 6 proposes a Big Data architecture for nowcasting social and economic variables, explaining its different modules; finally, Section 7 draws some concluding remarks.

2. Related work

Since the advent of the concept of “Big Data” two decades ago, some architectures to manage and analyze such data in different fields have been proposed, having their technical roots in distributed computing paradigms such as grid computing (Berman et al., 2003). However, the current data explosion also referred to as “Data Big Bang” (Pesenson et al., 2010) in which there is a daily generation of vast quantities of data from a variety of formats and sources, is revealing the fullest meaning of “Big Data”.

The particular properties and challenges that the current Big Data context opens require specific architectures for information systems particularly designed to retrieve, process, analyze and store such volume and variety of data. Therefore, we are living the constant births of new technologies conceived to be useful in this context such as, to mention some, cloud and exascale computing (Bahrami and Singhal, 2014; Reed and Dongarra, 2015). Given this recent technological and data revolution, research in this topic is in its early stage (Chen et al., 2014). In this section, we review the novel and incipient research works that develop general frameworks and specific architectures for adopting the Big Data approach in different fields from the point of view of data analytics applications.

Pääkkönen and Pakkala (2015) proposed a reference architecture for Big Data systems based on the analysis of some implementation cases. This work describes a number of functionalities expected to be considered when designing a Big Data architecture for a specific knowledge field, business or industrial process. These include: Data sources, data extraction, data loading and preprocessing, data processing, data analysis, data transformation, interfacing and visualization, data storage and model specification. Besides that, Assunção et al. (2015) reflected on some components that should be present in any Big Data architecture by depicting the four most common phases within a Big Data analytics workflow: Data sources, data management (including tasks such as preprocessing and filtering), modelling, and result analysis and visualization. This scheme was put in relation to cloud computing, whose potential and benefits for storing huge amounts of data and performing powerful calculus are positioning it as a desirable technology to be included in the design of a Big Data architecture. Concretely, the role of cloud computing as part of a Big Data system has been explored by Hashem et al. (2015).

About architectures for specific domains, Zhang et al. (2017) proposed a Big Data analytics architecture with the aim of exploiting industrial data to achieve cleaner production processes and optimize the product lifecycle management. This architecture works in four main stages: in stage 1, services of product lifecycle management, such as design improvement, are applied; in stage 2, the architecture acquires and integrates Big Data from different industrial sources, such as sensors; in stage 3, Big Data is processed and stored depending on their structure; finally, in stage 4, Big Data mining and knowledge discovery is conducted by means of four layers: the data layer (mixing data), the method layer (data extraction), the result layer (data mining) and the application layer (meeting the demands of the enterprise). Results from last stage fill the ERP systems and are used along with decision support systems to improve product-related services and give feedback in all product lifecycle stages.

In the domain of healthcare, a complete and specific Big Data analytics architecture was developed by Wang et al. (2016a). This architecture was based on the experiences about best practices in implementing Big Data systems in the industry, and was composed of five major layers: first, the data layer, which includes the data sources to be used for supporting operations and problem solving; second, the data aggregation layer, which is in charge of acquiring, transforming and storing data; third, the analytics layer, which is in charge of processing and analyzing data; fourth, the information exploration layer, which works by generating outputs for clinical decision support, such as real-time monitoring of potential medical risks; last, the data governance layer, which is in charge of managing business data throughout its entire lifecycle by applying the proper standards and policies of security and privacy. This layer is particularly necessary in this case given the sensibility of clinical data.

The review of these architectures evidenced some common modules or functionalities. After homogenizing the different names for modules very similar responsibilities, and considering their sequence in the process, they can be summarized as follows: first, a data module, which includes different sources of data with different formats; second, a data preprocessing module, which includes data extraction, integration and transformation; third, a data analytics module, which includes modelling and analysis techniques for knowledge discovery; and fourth, a results and visualization module, which includes tools for representing the results in a way useful for the firm or organization.

However, there are other functionalities whose location within the Big Data architecture is not homogeneous across the different proposals. For instance, the data storage responsibilities, which are basic for enabling data reuse and bringing access to previous results, have been included in a variety of places, ranging from being included in the data module (Assunção et al., 2015) or the preprocessing module (Wang et al., 2016a; Zhang et al., 2017), to being a macro-functionality present in each module of the architecture (Pääkkönen and Pakkala, 2015). The last approach is better reflecting the nature and complexity of Big Data analysis, given that not only the original data requires storage, but also