Data Science and Distributed Intelligence

Alfredo Cuzzocrea
ICAR-CNR and University of Calabria, Italy

Mohamed Medhat Gaber
Robert Gordon University, Aberdeen, UK

INTRODUCTION

The two terms *Big Data* (Stonebraker & Hong, 2012) and *MapReduce* (Dean & Ghemawat, 2008) have dominated the scene in the intelligent data analysis field during the last two years. They are in fact the cause and effect of the rapid growth in data observed in the digital world. The phenomenon of very large databases and very high rate streaming data has been coined recently as Big Data. The largest two databases for Amazon account for 42 terabytes of data in total, and YouTube receives at least 65,000 new videos per day. Such figures increase every day and people are literally drowning in high waves of data. Making sense out of this data has become more important than ever in the knowledge era. With the birth of *learning from data streams*, Mutukrishnan in his later published book (Muthukrishnan, 2005) has defined data streams as "*data arriving in a high rate that challenges our computation and communication capabilities*." In fact, this definition is now more true than back then. In spite of the continuous advances in our computation and communication capabilities, the data growth has been much faster, and the problem has become even more challenging. As a natural reaction to this worsening, a number of advanced techniques for data streams have been proposed, ranging from *compression paradigms* (e.g., (Cuzzocrea et al., 2004a, 2004b, 2005; Cuzzocrea & Chakravarthy, 2010)), mainly inherited by previous experiences in *OLAP data cube compression* (e.g., (Cuzzocrea, 2005; Cuzzocrea & Serafino, 2009)) to intelligent approaches that successfully exploit the nature of such data sources, like their *multidimensionality*, to gain in effectiveness and efficiency during the processing phases (e.g., (Cuzzocrea, 2009)), and recent initiatives that are capable of dealing with complex characteristics of such data sources, like their *uncertainty* and *imprecision*, as dictated by modern stream applicative settings (e.g., *social networks*, *Sensor Web*, *Clouds* – (Cuzzocrea, 2011)).

Addressing such challenges has kept Data Mining and Machine Learning practitioners and researchers busy with exploring the possible solutions. MapReduce has come as a potentially effective solution when dealing with large datasets, by enabling the breakdown of the main process into smaller tasks. Each of these tasks could be performed either in a *parallel* or *distributed* processing mode of operation. This allows the speed-up of performing complex data processing tasks, in an attempt to catch up with high speed large volume of data generated by scientific applications (Jiang et al., 2010), such as the promising contexts of *analytics over large-scale multidimensional data* (e.g., (Cuzzocrea et al., 2011)) and large-scale sensor network data processing (e.g., (Yu et al., 2012)). With Big Data and MapReduce at the front of the scene, a new term describing the process of dealing with very large dataset has been coined, *Data Science*.

In line with this, when these kind of dataset are processed on top of a service-oriented infrastructure like the novel *Cloud Computing* one (Agrawal et al., 2011), the terms "*Database as a Service*" (DaaS) (Hacigumus et al., 2002) and "*Infrastructure as a Service*" (IaaS) arise, and it is become critical to understand how Data Science can be coupled with distributed, service-oriented infrastructures, with novel and promising computational metaphors. Hence, due to the inherent distributed nature of computational infrastructures like Clouds (but also *Grids* (Foster et al., 2001)), it is natural to view *Distributed Intelligence* as the most natural underlying paradigm for novel Data Science challenges.

DOI: 10.4018/978-1-4666-5888-2.ch166

BACKGROUND

In his famous article "What is Data Science?" Loukides (2010) has enumerated differences between Data Science and traditional statistical analysis. Mainly, Data Science deals with the whole process of gathering data, pre-processing them and finally making sense out of them, producing what he termed as *data products*. This definition may be confused with any definition given to Data Mining and Data Warehousing processes. What really makes Data Science different, however, is the holistic approach when looking at producing a data product. This is especially true with the large volumes of noisy and unstructured data generated in our daily lives, from social media to search terms on Google. Traditional Data Mining and Warehousing strategies become no longer valid when dealing with such large and dynamic data sources.

Thus, the phenomenon of Big Data has dictated the emergence of a new field that encompasses a number of well-established areas, including at the front line, Data Mining and Warehousing. This is the Data Science field, a term that researches will encounter very often for some years to come. Scaling up the data analysis techniques to cope with Big Data has spotted the light on old functional programming functions, map and reduce, giving raise to the MapReduce computational paradigm. In the following subsections, a discussion of the Big Data phenomenon and how the two functions map and reduce have helped scaling up Big Data problems within the MapReduce paradigm is provided.

The Big Data Phenomenon

Soulellis (2012) has enumerated a number of examples of Big Data. These include:

- Approximately, one zettabyte (i.e., 1,000,000,000,000 bytes) of data have been produced in 2010.
- It is estimated that 8 zettabytes of data will be produced in 2015.
- More than 30,000 tweets are sent every minute actually.

All these examples well-describe the Big Data phenomenon that characterizes actual information systems. More interestingly and in addition to these examples, 90% of our data was the result of only the last two years of data production.

As a consequence, a big challenge with such huge data arise, and adequate analysis of these data can help advancing knowledge greatly. There is no doubt that there is a great business advantage when enterprises are able to use such data to guide their decision making. It is a well-known news story that *GAP* store chain management have reverted their decision to change the company's logo when sentiment extracted from social media revealed that the customers did not like the new logo (BCC, 2011). Another example is the controversial news story that *TARGET* department store have been able to predict that a teenage girl is pregnant using her new pattern of purchasing (Hill, 2012).

Not only business enterprises can benefit from such large data repositories, scientific discoveries can be also drawn from big data collected using advanced instruments generating data at very high rates. *Galaxy Zoo* (Lintott et al., 2008) is one example of large data repository that uses the emergence of citizen science. Citizen science uses crowd annotation and data collection for the use in scientific research. In *Galaxy Zoo*, a very large collection of images representing galaxies are provided for users to annotate. For instance, Figure 1 shows the Milky Way galaxy exploited for annotation purposes.

The MapReduce Computational Paradigm

MapReduce is a programming model that uses a *divide and conquer* method to speed-up processing large datasets (Dean & Ghemawat, 2010). It has been used in 2003 for the implementation of inverted index within the *Google Search Engine* in order to efficiently handle the search process. Also, it has been successfully exploited to handle large scale Machine Learning and Text Analytics tasks within *Google Analytics*. *Hadoop* (Apache, 2012) is the widely-known open source implementation of MapReduce.

The model of *MapReduce* has two main functions map and (reduce). The map function processes a key/value pair to produce a number of intermediate key/