Chapter 49
A Distributed and Scalable Solution for Applying Semantic Techniques to Big Data

Alba Amato
Second University of Naples, Aversa, Italy

Salvatore Venticinque
Second University of Naples, Aversa, Italy

Beniamino Di Martino
Second University of Naples, Aversa, Italy

ABSTRACT

The digital revolution changes the way culture and places could be lived. It allows users to interact with the environment creating an immense availability of data, which can be used to better understand the behavior of visitors, as well as to learn about their thoughts on what the visit creates excitement or disappointment. In this context, Big Data becomes immensely important, making possible to turn this amount of data in information, knowledge, and, ultimately, wisdom. This paper aims at modeling and designing a scalable solution that integrates semantic techniques with Cloud and Big Data technologies to deliver context aware services in the application domain of the cultural heritage. The authors started from a baseline framework that originally was not conceived to scale when huge workloads, related to big data, must be processed. They provide an original formulation of the problem and an original software architecture that fulfills both functional and not-functional requirements. The authors present the technological stack and the implementation of a proof of concept.

INTRODUCTION

The digital revolution changes the way culture and places could be lived. It allows users to interact with the environment creating an immense availability of data, which can be used to better understand the behavior of visitors, as well as to learn about their thoughts on what the visit creates excitement or disappointment. Supporting the visit of an archaeological site by handled devices allows for collecting a

DOI: 10.4018/978-1-4666-9840-6.ch049

lot of data, for example about the movements of those who visited the exhibition, about which artifacts they focused on, which ones has avoided seeing, the search performed, the feedback submitted, etc. Additional information can be collected from various sources such as social networks, data warehouse, web applications, networked machines, virtual machines, sensors over the network, etc. It is necessary to think about how and where to processing them. It is necessary a scalable, distributed storage systems, a set of flexible data models that allow for an effective utilization of available technologies and computational resources.

The need to store, manage, and treat the ever increasing amounts of data is becoming increasingly felt. The effort spent in redesigning and optimizing data storage for analysis requests could result in poor performance. In fact current databases and management tools are inadequate to handle complexity, scale, dynamism, heterogeneity, and growth of such systems. Big data technologies can address the problems related to the collection of data streams of higher velocity and higher variety.

Big Data are an important and valuable resource for innovation, competition and productivity if properly managed. Gartner defines Big Data as ”high volume, velocity and/or variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation” (Gartner, 2012). Those data set are enormous; their size is beyond the ability of systems of typical database to capture, integrate, manage and analyze them. But the huge size is not the only property of Big Data. Only if the information has the characteristics of Volume, Velocity and/or Variety we can talk about Big Data (P. Zikopoulos, and C. Eaton, 2011) Volume refers to the fact that we are dealing with ever-growing data expanding beyond terabytes into petabytes, and even exabytes (1 million terabytes). Variety refers to the fact that Big Data is characterized by data that often come from heterogeneous sources such as machines, sensors and unrefined ones, making the management much more complex. Finally, the third characteristic, that is velocity that, according to Gartner (Gartner, 2011) “means both how fast data is being produced and how fast the data must be processed to meet demand”. In fact in a very short time the data can become obsolete. IBM (M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, 2012) proposes the inclusion of veracity as the fourth Big Data attribute to emphasize the importance of addressing and managing the uncertainty of some types of data. With the amount of data being produced every day, there is the need to unlock the unnamed fifth V of big data: VALUE. According to analysts with Forrester (Forrester, 2014), most organizations today use less than 5% of the data that is available to them. As our capability to collect data has increased, our ability to store, sort and analyze it has diminished. In this context, Big Data becomes immensely important, making possible to turn into this amount of data in information, knowledge, and, ultimately, wisdom. The requirements of many applications are changing and require the adoption of these technologies. NoSQL databases ensure better performance than RDBMS systems in various use cases, most notably those involving big data. But the choice of the one that best fits the application requirements is a challenge for the programmers that decide to develop a scalable application. There are many differences among the available products and also among the level of maturation on them. From a solution point of view it is necessary a clear analysis of the application context. In particular we focused on technologies that operate in pervasive environments, which can benefit from the huge information available but need to be rethought to extract knowledge and improve the context awareness in order to customize the services.

Even if a storage solution conceived ad hoc to provide good performance and to scale, exploiting the elasticity of new computing paradigms like the Cloud, the design and programming of processing functions must follow an effective methodology. In particular the utilization of semantic techniques for