A Novel Multidimensional Approach to Integrate Big Data in Business Intelligence

Alejandro Maté, Department of Software and Computing Systems, University of Alicante, Alicante, Spain
Hector Llorens, Department of Software and Computing Systems, University of Alicante, Alicante, Spain
Elisa de Gregorio, Department of Software and Computing Systems, University of Alicante, Alicante, Spain
Roberto Tardío, Department of Software and Computing Systems, University of Alicante, Alicante, Spain
David Gil, Department of Computing Technology and Data Processing, University of Alicante, Alicante, Spain
Rafa Muñoz-Terol, Department of Software and Computing Systems, University of Alicante, Alicante, Spain
Juan Trujillo, Department of Software and Computing Systems, University of Alicante, Alicante, Spain

ABSTRACT

The huge amount of information available and its heterogeneity has surpassed the capacity of current data management technologies. Dealing with huge amounts of structured and unstructured data, often referred as Big Data, is a hot research topic and a technological challenge. In this paper, the authors present an approach aimed to enable OLAP queries over different, heterogeneous, data sources. Their approach is based on a MapReduce paradigm, which integrates different formats into the recent RDF Data Cube format. The benefits of their approach are that it is capable of querying different sources of information, while maintaining at the same time, an integrated, comprehensive view of the data available. The paper discusses the advantages and disadvantages, as well as the implementation challenges that such approach presents. Furthermore, the approach is evaluated in detail by means of a case study.

Keywords: Big Data, Business Intelligence, Conceptual models, OLAP cube, RDF/XML

DOI: 10.4018/JDM.2015040102
INTRODUCTION

Nowadays, more and more information present on Internet is being considered for aiding in decision making. However, the size of the information produced per year already exceeds the zettabytes (Zikopoulos et al., 2011). It is already not possible to process, store, and manage all this information in local servers even for the systems of the biggest Business Intelligence (BI) companies. The possibility of making on-line analytical processing (OLAP) queries over high amounts of information, while being able to retrieve only the relevant information at each moment, would provide important benefits. However, given the heterogeneity and size of the data used, the effort required to harness the power of all this information can not be afforded by individual companies. This phenomenon is often referred to as Big Data (Cohen et al., 2009).

Until the recent years, not much structured information was available on the Internet. Most of this information was textual information written in the most widely spoken natural languages. Structured data was mainly stored in private databases, being accessible only by their owners. Nowadays, another breaking phenomena, Open-Data (Lindman et al., 2013), is changing this situation drastically. In the same manner as Wikipedia freely brings unlimited access to lots of semi-structured information, many institutions and communities have decided to publish and share on the Internet the information they manage. For example, the governments of some countries have decided to publish their information in order to increase their transparency (e.g. data.gov.com). Following this trend, many other kinds of data, such as road traffic, are also becoming open, thus increasing the number of available sources of data. Some studies (Colomo-Palacios et al., 2012) deep into the wide possibilities of these new data sources.

In this context, imagine the possibility of performing OLAP analysis over a distributed model mixing private local data with the open linked data available. A company with a private database (sales, customers, strategy) could benefit from querying this data together with the linked data, which can provide a significant enhancement for the company OLAP capabilities. The following are just few examples of what could be done with such a model: “Which countries are suffering decreases in sales and GDP drops” (WorldBank stats), “Which of our products decreasing in sales in the last quarter have negative opinions on Twitter?”.

This kind of queries would clearly provide a good support to improve the decision making process in companies. The difficulty behind this challenge is mainly focused on the heterogeneity of the sources from which the information would be extracted, as well as on the importance of an efficient distributed model to make these queries computationally viable.

The aim of this paper is to propose and analyze an integrated approach to enable OLAP queries over heterogeneous data sources, where each data source may contain different internal and external dimensions. We propose an approach based on the MapReduce strategy (Dean and Ghemawat, 2008) capable of dividing a query, and distributing it to different nodes that access different datasets in a variety of formats. The output of these nodes is then seamlessly integrated, making the process transparent for the user. In order to increase the extensibility of our proposal, we base our approach in standards, using SPARQL as the query language for the distributing/integrating module and as input for the nodes.

The remainder of this paper is structured as follows. In the next section, the related work is reviewed. Thereafter, our proposal of an integrated model to access Big Data is presented. Finally, in the last two sections, we include a discussion of the advantages of the model and the difficulties related to its implementation as well as the further directions of this research.
Fabric Database and Fuzzy Logic Models for Evaluating Fabric Performance
www.igi-global.com/chapter/fabric-database-fuzzy-logic-models/20367?camid=4v1a

Dimensions of UML Diagram Use: A Survey of Practitioners
www.igi-global.com/article/dimensions-uml-diagram-use/3379?camid=4v1a