Chapter 23
Big Data Warehouse Automatic Design Methodology

Francesco Di Tria
Università degli Studi di Bari Aldo Moro, Italy

Ezio Lefons
Università degli Studi di Bari Aldo Moro, Italy

Filippo Tangorra
Università degli Studi di Bari Aldo Moro, Italy

ABSTRACT

Traditional data warehouse design methodologies are based on two opposite approaches. The one is data oriented and aims to realize the data warehouse mainly through a reengineering process of the well-structured data sources solely, while minimizing the involvement of end users. The other is requirement oriented and aims to realize the data warehouse only on the basis of business goals expressed by end users, with no regard to the information obtainable from data sources. Since these approaches are not able to address the problems that arise when dealing with big data, the necessity to adopt hybrid methodologies, which allow the definition of multidimensional schemas by considering user requirements and reconciling them against non-structured data sources, has emerged. As a counterpart, hybrid methodologies may require a more complex design process. For this reason, the current research is devoted to introducing automatisms in order to reduce the design efforts and to support the designer in the big data warehouse creation. In this chapter, the authors present a methodology based on a hybrid approach that adopts a graph-based multidimensional model. In order to automate the whole design process, the methodology has been implemented using logical programming.

1. INTRODUCTION

Big data warehousing refers commonly to the activity of collecting, integrating, and storing (very extra) large volumes of data coming from data sources, which may contain both structured and unstructured data. However, volume alone does not imply big data. Further and specific issues are related to the velocity in generating data, and their variety and complexity.

DOI: 10.4018/978-1-4666-9840-6.ch023

The increasing volume of data stored in data warehouses is mainly due to their nature of preserving historical data, for performing statistical analyses and extracting significant information, hidden relationships, and regular patterns from data. Other factors that affect the size growth derive from the necessity of integrating several data sources, each of them provides a different variety of data that contribute to enrich the types of analyses, by correlating a large set of parameters. Furthermore, some data sources—such as Internet transactions, networked devices and sensors, for example—generate billions of data very quickly. These data should update the data warehouse as soon as possible, in order to gain fresh information and make timely decisions (Helfert & Von Maur, 2001).

These issues affect the design process, because big data warehouses must integrate heterogeneous data to be used to perform analyses that consider many points of view, and to produce complex schemas having cubes with high number of dimensions. Furthermore, they must be capable of quickly integrating new data sources through a minimal data modelling process.

To summarize this, new aspects for data warehouses supporting analyses of Big Data have been stated in Cohen *et al.* (2009). Big data warehouses have to be (i) *magnetic* for they must attract all the data sources available in an organization; (ii) *agile* for they should support continuous and rapid evolution; and (iii) *deep* in that they must support analyses more sophisticated that traditional OLAP functions.

1.1. Background Approaches to Automatic Design

In the mentioned scenario, traditional design methodologies, which are based on two opposite approaches—data-oriented and requirement-oriented— (Romero & Abelló, 2009), are not able to solve problems when facing big data.

In fact, methodologies adopting a data-oriented approach are devoted to define multidimensional schemas on the basis of the remodelling of the data sources. These data must be strongly structured, since functional dependencies are taken into account in the remodelling phase (dell'Aquila *et al.*, 2009). Then, these methodologies are not able to create a multidimensional schema from non-structured data sources. Furthermore, in presence of a high number of data sources, the process of solving semantic and syntactical inconsistencies among the different databases can be a very hard task without using an ontological approach. This reengineering process is individually executed by the designer who minimizes the involvement of end users and, consequently, goes towards a possible failure of their expectations. In the worst case, the data warehouse is completely useless and the design process must be revised.

On the opposite side, methodologies adopting a requirement-oriented approach define multidimensional schemas using business goals resulting from the decision makers' needs. The data sources are considered later, when the Extraction, Transformation, and Loading (ETL) phase is addressed. In the feeding plan, concepts of the complex multidimensional schema (such as facts, dimensions, and measures) have to be mapped on the data sources, in order to define the procedures to populate the data warehouse by cleaned data. At this point, the definition of these procedures can be very difficult and, in the worst case, it may happen that the designer discovers that the needed data are not currently available at the sources. On the other hand, some data sources containing interesting information, albeit available, may have been omitted or not exploited.

However, each of these two approaches has valuable advantages. So, the necessity emerged to adopt a hybrid methodology which takes into account their best features (Di Tria *et al.*, 2012; Di Tria *et al.*, 2011; Mazón & Trujillo, 2009; Mazón *et al.*, 2007; Giorgini *et al.*, 2008; Bonifati *et al.*, 2001). As a