Chapter 1
From User Requirements to Conceptual Design in Data Warehouse Design
A Survey

Matteo Golfarelli
DEIS - University of Bologna, Italy

ABSTRACT
Conceptual design and requirement analysis are two of the key steps within the data warehouse design process. They are to a great extent responsible for the success of a data warehouse project since, during these two phases, the expressivity of the multidimensional schemata is completely defined. This chapter proposes a survey of the literature related to these design steps and points out pros and cons of the different techniques in order to help the reader to identify crucial choices and possible solutions more consciously. Particular attention will be devoted to emphasizing the relationships between the two steps describing how they can be jointly used fruitfully.

INTRODUCTION
Data Warehouse (DW) systems are used by decision makers to analyze the status and the development of an organization. DWs are based on large amounts of data integrated from heterogeneous sources into multidimensional schemata which are optimized for data access in a way that comes natural to human analysts. Generally speaking, a multidimensional schema is made up of facts, measures and dimensions. Facts are a focus of interest for the decision-making process (e.g. sales, orders) and can be monitored through measures and dimensions. Measures are numerical KPIs (e.g., quantity of product sold, price of the products, etc.), and dimensions represent the context for analyzing these measures (e.g., time, customer, product, etc.). Owing to their specificities, the development of DWs is particularly complex and requires ad-hoc methodologies and an appropriate life-cycle.

Conceptual design and requirement analysis are two of the key steps within the DW design process. While they were partially neglected in the first era of data warehousing, they have received greater attention in the last ten years.
The research literature has proposed several original approaches for conceptual modeling in the DW area, some based on extensions of known conceptual formalisms (e.g. E/R, UML), some based on ad hoc ones. Remarkably, a comparison of the different models pointed out that, abstracting from their graphical form, the core expressivity is similar, thus proving that the academic community has reached an informal agreement on the required expressivity.

On the other hand, the proposed solutions are not always coupled with an appropriate technique for requirement analysis to form a methodological approach ensuring that the resulting database will be well-documented and will fully satisfy the user requirements. DW specificities make these two steps even more related than in traditional database systems; in fact the lack of settled user requirements and the existence of operational data sources that fix the set of available information make it hard to develop appropriate multidimensional schemata that, on the one hand, fulfill user requirements and on the other, can be fed from the operational data sources.

This paper proposes a survey of the literature related to these design steps in order to help the reader make crucial choices more consciously. In particular, after a brief description of the DW lifecycle, the specific problems arising during requirement analysis and conceptual design are presented. The approaches to requirement analysis are then surveyed and their strengths and weaknesses are discussed. Afterwards, the literature related to the DW conceptual models is also surveyed and the core expressivity of these models is discussed in order to enable the reader to understand which are the relevant pieces of information to be captured during user-requirements analysis.

**BACKGROUND**

The DW is acknowledged as one of the most complex information system modules and its design and maintenance is characterized by several complexity factors that determined, in the early stages of this discipline, a high percentage of real project failures. A clear classification of the critical factors of data warehousing projects was already available in 1997 when three different risk categories were identified (Demarest, 1997), namely socio-technical i.e. related to the impact a DW has on the decisional processes and political equilibriums, technological i.e. related to the usage of new and continuously evolving technologies, and design-related i.e. related to the peculiarities of this kind of systems. The awareness of the critical nature of the problems and the experience accumulated by practitioners determined the development of different design methodologies and the adoption of proper life-cycles that can increase the probability of completing the project and fulfill the user requirements.

The choice of a correct life-cycle for the DW must take into account the specificities of this kind of system, that according to Giorgini et al. (2007), are summarized as follows:

a) DWs rely on operational databases that represent the sources of the data.

b) User requirements are difficult to collect and usually change during the project.

c) DW projects are usually huge projects: the average time for their construction is 12 to 36 months and their average cost ranges from 0.5 to 10 million dollars.

d) Managers are demanding users that require reliable results in a time compatible with business needs.

While there is no consensus on how to address points (a) and (b), the DW community has agreed on an approach that cuts down costs and time to make a satisfactory solution available to the final users. Instead of approaching the DW development as a whole in a top-down fashion, it is more convenient to build it bottom-up working on single data marts (Jensen et al., 2004). A *data mart* is
Related Content

Off-Line Signature Recognition
[www.irma-international.org/chapter/off-line-signature-recognition/10719/](www.irma-international.org/chapter/off-line-signature-recognition/10719/)

Peer-to-Peer Data Clustering in Self-Organizing Sensor Networks
Stefano Lodi, Gabriele Monti, Gianluca Moro and Claudio Sartori (2010). *Intelligent Techniques for Warehousing and Mining Sensor Network Data* (pp. 179-212).
[www.irma-international.org/chapter/peer-peer-data-clustering-self/39546/](www.irma-international.org/chapter/peer-peer-data-clustering-self/39546/)

Mosaic-Based Relevance Feedback for Image Retrieval
[www.irma-international.org/chapter/mosaic-based-relevance-feedback-image/10713/](www.irma-international.org/chapter/mosaic-based-relevance-feedback-image/10713/)

Graph-Based Data Mining
[www.irma-international.org/chapter/graph-based-data-mining/10934/](www.irma-international.org/chapter/graph-based-data-mining/10934/)

Cluster-Based Input Selection for Transparant Fuzzy Modeling
[www.irma-international.org/chapter/cluster-based-input-selection-transparant/7691/](www.irma-international.org/chapter/cluster-based-input-selection-transparant/7691/)