Research in data warehouse modeling and design: Dead or alive?

Conference Paper · January 2006

DOI: 10.1145/1183512.1183515 · Source: DBLP

4 authors:

Stefano Rizzi
University of Bologna
179 PUBLICATIONS **4,466** CITATIONS
SEE PROFILE

Alberto Abelló
Universitat Politècnica de Catalunya
121 PUBLICATIONS **1,900** CITATIONS
SEE PROFILE

Jens Lechtenbörger
University of Münster
51 PUBLICATIONS **1,398** CITATIONS
SEE PROFILE

Juan Trujillo
University of Alicante
277 PUBLICATIONS **4,459** CITATIONS
SEE PROFILE

Research in Data Warehouse Modeling and Design: Dead or Alive?

Stefano Rizzi
University of Bologna - Italy
srizzi@deis.unibo.it

Alberto Abelló
Polytechnical University of Catalunya - Spain
aabello@lsi.upc.edu

Jens Lechtenbörger
University of Münster - Germany
lechten@wi.uni-muenster.de

Juan Trujillo
University of Alicante - Spain
jtrujillo@dlsi.ua.es

ABSTRACT

Multidimensional modeling requires specialized design techniques. Though a lot has been written about how a data warehouse should be designed, there is no consensus on a design method yet. This paper follows from a wide discussion that took place in Dagstuhl, during the Perspectives Workshop "Data Warehousing at the Crossroads", and is aimed at outlining some open issues in modeling and design of data warehouses. More precisely, issues regarding conceptual models, logical models, methods for design, interoperability, and design for new architectures and applications are considered.

Categories and Subject Descriptors

H.4.2 [**Information Systems Applications**]: Types of Systems—*Decision support*; H.2.1 [**Database Management**]: Logical Design

General Terms

Design

Keywords

Data warehouse design, multidimensional modeling

1. INTRODUCTION

It is well known that data warehouses (DWs) are focused on decision support rather than on transaction support, and that they are prevalently characterized by an OLAP workload. Traditionally, OLAP applications are based on multidimensional modeling, that intuitively represents data under the metaphor of a cube whose cells store events that occurred in the business domain. Adopting the multidimensional model for DWs has a two-fold benefit. On the one hand, it is close to the way of thinking of data analyzers and, therefore, it helps users understand data; on the other hand, it supports performance improvement as its simple structure allows designers to predict users' intentions.

Multidimensional modeling and non-OLTP workloads require specialized design techniques. The most cited difference between design for transactional databases and DWs is denormalization, yet DW design has several other relevant peculiarities. Though a lot has been written about how a DW should be designed, there is no consensus on a design method yet. Most methods agree on the opportunity for distinguishing between a phase of *conceptual design* and one of *logical design*, like in [24, 31, 45]. Conceptual design aims at deriving an implementation-independent and expressive conceptual schema for the DW, according to the chosen conceptual model, starting from the user requirements and from the structure of the source databases. Logical design takes the conceptual schema and creates a corresponding logical schema on the chosen platform by considering some set of constraints (e.g., concerning disk space or query answering time). Several methods (e.g., [24]) also support a phase of *physical design*, that addresses all the issues specifically related to the suite of tools chosen for implementation – such as indexing and allocation. In some cases, a phase of *requirement analysis* (e.g., [19]) is separately considered. From the functional point of view, the relationships between these phases can be summarized as in Figure 1 (in practice, this process will likely include feedback loops that allow to re-enter previous phases). Unfortunately, though most vendors of DW technology propose their own CASE solutions (that very often are just wizards capable of supporting the designer during the most tedious and repetitive phases of design), the only tools that currently promise to effectively automate some phases of design are just research prototypes (see for instance [25, 75]).

This paper arises as an afterthought following a wide discussion that took place in Dagstuhl, during the Perspectives Workshop "Data Warehousing at the Crossroads" (August 2004). While the aim of the seminar was to discuss the current trends in data warehousing and to pave the way for future research in the whole field, here we will specifically focus on modeling and design, trying to answer the following question: "Has research on this topic come to an end? If not, what's left to do?" Thus, in this paper, benefiting from the fruitful discussions that took place there between

DOLAP'06, November 10, 2006, Arlington, Virginia, USA.

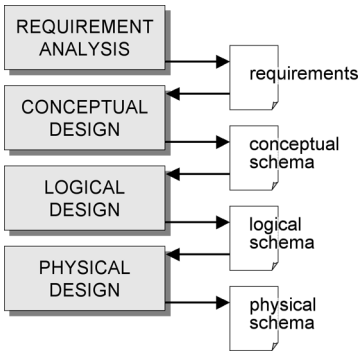

Figure 1: The core phases in DW design

all participants, we survey some topics related to DW modeling and design and outline the issues that, in our view, still need further exploration. More precisely, in Sections 2, 3, 4, 5, and 6 we address, respectively, conceptual models, logical models, methods for design, interoperability, and design for new architectures and applications.

2. CONCEPTUAL MODELING

Conceptual modeling provides a high level of abstraction in describing the warehousing process and architecture in all its aspects, aimed at achieving independence of implementation issues. Conceptual modeling is widely recognized to be the necessary foundation for building a database that is well-documented and fully satisfies the user requirements; usually, it relies on a graphical notation that facilitates writing, understanding, and managing conceptual schemata by both designers and users.

In the literature, conceptual modeling for DWs has been tackled from mainly two points of view so far:

- *Multidimensional modeling.* The existing approaches may be framed into three categories: extensions to the Entity-Relationship model (e.g, [18, 66]), extensions to UML (e.g., [1, 46]), and ad hoc models (e.g., [22, 31]). While all models have the same core expressivity, in that they all allow the basic concepts of the multidimensional model to be represented, they significantly differ as to the possibility of representing more advanced concepts such as irregular hierarchies, many-to-many associations, and additivity.
- *Modeling of ETL.* The focus is to model the ETL process either from the functional [79], the dynamic [6], or the static [10] point of view. Though the research on ETL modeling is probably less mature than that on multidimensional modeling, we believe that it will have a very relevant impact on improving the overall reliability of the design process and on reducing its duration.

While apparently a lot of work has been done in the field of conceptual modeling, we believe that some very important issues still remain open, as detailed in the following.

2.1 Lack of a standard

Though several conceptual models have been proposed, none of them has been accepted as a standard so far, and all vendors propose their own proprietary design methods. The main reasons for this, we argue, can be summarized as follows: (i) there is still no agreement from both the research and industrial communities about which are the most relevant multidimensional properties to be modeled; (ii) although the conceptual models devised are semantically rich, some of the modeled properties cannot be expressed in the target logical models, so the translation from conceptual to logical is incomplete (see Section 3.1); and (iii) commercial CASE tools currently enable designers to directly draw logical schemata, thus no industrial push is given to any of the models. On the other hand, we believe that a unified conceptual model for DWs, implemented within sophisticated CASE tools, would be a valuable support for both the research and industrial communities. It should be formally well-founded, but at the same time easily usable and understandable by designers. It should support integrated modeling of the DW architecture, deployment, sources, mappings, ETL, facts, workloads, etc. Finally, it should be expressive and flexible enough not only to enable representation of requirements coming from the classical enterprise domains, but also to support the peculiar issues and constraints arising in unusual and emerging domains and applications (such as those based on streaming data or geographical information). Notice the difficulties on defining such model, due to the clear antagonism between expressiveness and understandability.

2.2 Modeling security

Information security is a basic requirement for a wide range of applications. In the case of DWs, among the different aspects of security, confidentiality (i.e., ensuring that users can only access the information they have privileges for) is particularly relevant, because business information is very sensitive and can be discovered by executing a simple query. Unfortunately, the classical security model used in transactional databases – centered on tables, rows, and attributes – is unsuitable for DWs. For instance, two queries obtained one from another through a simple drill-down operation (thus differing only in their aggregation levels) may involve the same table, rows, and columns, though the one formulated at the finest aggregation might reveal undesired details of data to the user. Thus, the classical security model should be replaced with an appropriate model centered on the main concepts of multidimensional modeling – such as facts, dimensions, and measures – and tightly integrated with the conceptual model adopted. In addition, as commonly recommended in software engineering, information security should be considered not in isolation but during all stages of the development life-cycle, from requirement analysis to implementation and maintenance.

Though most conceptual models for DWs in the literature do not address security, lately some interesting proposals were devised which define specific authorization and security models (e.g., [39, 32]). However, these proposals mainly deal with OLAP operations accomplished with OLAP tools, thus they are unsuitable for integration in multidimensional modeling as part of DW design. To the best of our knowledge, only two works consider security measures as integrated within conceptual modeling: the extended ADAPTed UML [61] and the UML extension presented in [17]. Although both approaches consider security from the early stages of a DW project, they can be considered preliminary works still requiring further research.

Thus, there are still a number of issues that should be tackled in security modeling:

- Devise a reliable and flexible security model that comprehensively considers all the components of the warehousing architecture, including ETL and data sources;
- Provide a method for transforming security models from the conceptual level into the logical level, and then into concrete implementations in target commercial platforms;
- Represent a complete and integrated hierarchy of roles and compartments for different groups of users, supported by a formal language to solve conflicts between different authorization rules.

2.3 Mining-aware design

Vendors like IBM and Microsoft already mix OLAP and data mining in their commercial tools. Nevertheless, with the notable exceptions of Han's *OLAM* [28] and, more recently, of *prediction cubes* [14], the research community in general and DW researchers in particular have not considered integrating OLAP and data mining as a hot topic. So far, DW design has been mainly targeted at designing OLAP cubes, and no attention has been paid to consider mining requirements from the early stages of design. Conversely, we believe that devising mining-aware design techniques and models raises a number of interesting research issues:

- How could mining results be gracefully incorporated into DWs? While some approaches to model mining patterns as first-class citizens in databases have been tempted [63], the only approaches to multidimensional modeling of patterns we are aware of are the work presented in [86], that incorporates the definition of association rules in the specification of conceptual schemata for DWs, and the prediction cubes proposed in [14], that support OLAP navigation of cells summarizing prediction models.
- How could DW and OLAP storage techniques support data mining algorithms by facilitating them in accessing large volumes of cleansed and integrated data? As suggested in [13], this may strongly enhance the scalability of data analysis.
- How would the two analysis techniques complement one another? Some suggestions in this direction can be found in [15, 37].

3. LOGICAL MODELING

Once the conceptual modeling phase is completed, the overall task of logical modeling is the transformation of conceptual schemata into logical schemata that can be optimized for and implemented on a chosen target system.

Considerable progress has been made in the area of multidimensional modeling, where target database systems are typically either relational or multidimensional. In relational implementations, the so-called star, constellation, and snowflake schemata are widely accepted to manage data cubes and are supported by various vendors. Concerning multidimensional implementations, several efficient multidimensional data structures such as condensed cubes [82, 16], dwarfs [72, 73], and QC-Trees [40] have been proposed to manage data cubes.

Nevertheless, we believe that some relevant challenges remain for future research, as summarized in the following subsections.

3.1 Semantic gap

With respect to fact modeling, there still is a semantic gap between advanced conceptual data models and relational or multidimensional implementations of data cubes. For instance, no commercial solutions can cope with generalization/specialization relationships in OLAP hierarchies [48]. Additionally, it appears to be an open problem how to represent dimension constraints [30] or even less expressive context dependencies [42], both of which explain the existence of null values in dimensions in logical implementations and allow to reason about summarizability with respect to *sets* of attributes. Moreover, a systematic treatment of summarizability addressing general aggregate functions beyond SUM remains an open issue [29]. Consequently, future research is necessary to bridge this semantic gap, i.e., to preserve all information captured by advanced conceptual multidimensional models in logical implementations. To this end, research could either investigate how to enrich meta-data for tool support in a systematic way or, more ideally, look for more expressive logical models while preserving good query performance. Clearly, without the support of more expressive logical models we cannot expect to achieve a streamlined design process that guarantees quality criteria (e.g., avoidance of inconsistent queries, control over null values, reduction of sparsity [42, 55]) to be satisfied and seriously takes security into account.

3.2 ETL modeling

The transformation of conceptual ETL schemata into logical ones as well as their optimization are not very well understood. Indeed, while [71, 78] present first steps towards the modeling and optimization of ETL processes at the logical level, [70] appears to be the only design method that includes an algorithmic transformation of conceptual into logical models.

Moreover, research on DW self-maintainability and independence (see, e.g., [62, 41]) has shown how to set up DWs in such a way that the maintenance processes can be simplified and made more efficient by avoiding maintenance queries. However, a combination of these results with ETL modeling techniques is still missing.

4. METHODS FOR DESIGN

While in the subsections above we have discussed the problems related to conceptual and logical models, in this subsection we are concerned with the techniques for building conceptual and logical schemata according to such models, considering them in the context of a comprehensive design framework that complies with good-design principles such as reusability, extendibility, and manageability.

Several techniques for automating single phases of DW design have been proposed in the literature (for instance, [22] for conceptual design, [74] for logical design, [27] for physical design, [78] for designing the ETL process). On the other hand, despite the basic role played by a well-structured methodological framework in ensuring that the DW designed fully meets the user expectations, a very few

comprehensive design methods have been devised so far (e.g., [12, 19, 24, 45]). Overall, we believe that some specific issues in design, discussed in the following subsections, have not been properly investigated yet. Besides, more generally, mechanisms should appear to coordinate all DW design phases allowing the analysis, control, and traceability of data and metadata along the project life-cycle. An interesting approach in this direction consists in applying the Model Driven Architecture in order to automate the inter-schema transformations from requirement analysis to implementation [50].

4.1 Requirements Analysis

Requirement analysis plays a key role within any software project to reduce the risk of failure. Nevertheless requirement analysis for DWs has not been given much attention so far, and it is often overlooked in DW projects mainly since (1) warehousing projects are long-term ones, in which most requirements cannot be stated from the beginning; and (2) requirements are poorly shared across organizations, unstable in time, and refer to information that must be derived from data sources [83].

The approaches to DW design are usually classified in two categories [83]. *Data-driven* approaches design the DW starting from a detailed analysis of the data sources; user requirements impact on design by allowing the designer to select which chunks of data are relevant for decision making and by determining their structuring according to the multidimensional model [31, 22]. *Requirement-driven* approaches start from determining the information requirements of end users, and how to map these requirements onto the available data sources is investigated only *a posteriori* [60, 51]. Some other authors, like [20, 11], use some kind of mixture of these two approaches, and consider both (i.e. availability of data and user requirements) at the same time, which appears to be a promising direction of research that is superior to isolated data-driven and requirement-driven approaches. Finally, a novel approach [36] is based on the definition of a set of *design patterns*, so that, once the needed pattern is found, it just has to be adapted to the available data and user requirements.

Though the approaches devised are promising, we believe that some further work needs to be done in order to provide designers with more usable and effective techniques for collecting information needs and quality-of-service requirements, and for translating them into (at least domain-specific, ideally general) conceptual models based on a common vocabulary between IT staff and decision makers. Thus, how quality-of-service can drive the design of the DW should be deeply studied.

4.2 Schema evolution

As several mature implementations of data warehousing systems are fully operational within medium to large contexts, the continuous evolution of the application domains is bringing to the forefront the dynamic aspects related to describing how the information stored in the DW changes over time. As concerns changes in data values, a number of approaches have been devised, and some commercial systems allow to track changes and to effectively query cubes based on different temporal scenarios [65]. Conversely, the problem of managing *changes on the schema level* (that may be demanded by changes either in the business domain or in the user requirements or in the sources) has only partially been explored, and no dedicated commercial tools or restructuring methods are available to the designer yet.

The approaches to management of schema changes in DWs can be framed into two categories, namely *evolution* [5, 76] and *versioning* [3, 21]: while both categories support schema changes, only the latter keeps track of previous versions. If one is sure that previous schema information will never be useful again, schema evolution offers adequate functionality. Otherwise (e.g., to guarantee consistent re-execution of old reports), schema versioning offers the strictly more powerful approach. Actually, in some versioning approaches, besides "real" versions determined by changes in the application domain, also "alternative" versions to be used for what-if analysis are considered [3]. Overall, we believe that versioning is better suited to support the complex analysis requirements of DW users as well as the DW characteristic of non-volatility. Thus, the main research challenges in this field are to provide effective versioning and data migration mechanisms, capable of supporting flexible queries that span multiple versions.

Considering the complexity of the ETL procedures, another very relevant issue is to devise techniques for propagating changes occurred in the source schemata to the ETL process. The obvious benefit in achieving these goals will be to keep the DW in sync with the business requirements, thus avoiding its obsolescence.

4.3 Quality metrics

Due to the strategic importance of DWs, it is absolutely crucial to guarantee their quality from the early stages of a project. While some relevant work on the *quality of data* has been carried out (e.g., [33, 34]), there is still no agreement on the *quality of the design process* and its impact on decision making. The most significant approaches to measuring the design quality can be framed as follows:

- *At the conceptual level.* There have been preliminary attempts towards defining metrics that allow the intuitive notions of quality of conceptual schemata to be replaced with quantitative measures (such as the number of facts, the number of degenerated dimensions, the number of shared hierarchy levels, etc.), in order to reduce subjectivity in evaluation and guide designers in their work [68, 77]. Obviously, the existence of a standard conceptual model could give a strong push in this direction.
- *At the logical/physical level.* Besides the recommendations and subjective criteria stated for instance in [38, 44], some works were focused on quantitatively evaluating the complexity of dimensional models [9]. Other relevant research directions include normal forms for DW [43, 42] and quality-driven view selection [7].

Overall, we believe it is necessary to devise more comprehensive metrics for measuring quality, encompassing both schema quality (e.g., to better model application requirements and to guarantee good querying performance) and data quality (e.g., to ensure timeliness of information and to take care of data aging). After their formal and empirical validation, these metrics will support the designer in evaluating and ranking different design alternatives; besides, they will be useful to better plan the project and meet user requirements, e.g., by predicting the cost and complexity of

later stages in design. Particular care should be taken in addressing the *traceability* of metrics, i.e., how metrics are translated from one phase of design to the next one, and in defining thresholds to discriminate "good" schemata from "bad" ones. Besides, techniques will be needed to *monitor* the metrics and appropriately respond to their deviations during the DW lifetime, in order to better manage extensions and evolutions. Finally, these metrics must be considered from the user point of view, by studying their impact on information analysis: methods must be devised to propagate data quality metrics to query results, like in [49], and to have data retrieval driven by the quality requirements expressed by users, like in [69].

5. INTEROPERABILITY AND METADATA

The heterogeneity in conceptual and logical models proposed for DWs, together with the wide variety of tools and software products available on the market, has lead to a broad diversity in metadata modeling. In practice, tools with dissimilar metadata are integrated by building complex metadata bridges, but some information is lost when translating from one form of metadata to another. Thus, there is a need for a standard definition of metadata in order to better support DW interoperability and integration, which is particularly relevant in the recurrent case of mergers and acquisitions.

Two industry standards developed by multi-vendor organizations have arisen in this context: the Open Information Model (OIM) [52] by the Meta Data Coalition (MDC) and the Common Warehouse Metamodel (CWM) [56] by the OMG (see [80] for a comparison of the two competing specifications). In 2000, MDC joined OMG for developing the CWM as a standard metadata model. The CWM is a platform-independent metamodel definition for interchanging DW specifications between different platforms and tools. It is based on the standards UML, XMI, and MOF, and basically provides a set of metamodels that are comprehensive enough to model an entire DW including data sources, ETL, multidimensional cubes, relational implementations, and so on. These metamodels are meant to be generic, external representations of shared metadata and to provide a framework for data exchange. Unfortunately, their expressivity is not sufficient to capture all the complex semantics represented by conceptual models, so they hardly can be used for effective integration of different DWs.

An alternative approach in this direction is described in [8], where a notion of dimension compatibility based on information consistency is proposed, aimed at cross-querying over autonomous, federated data marts. We believe that another interesting possibility for integration would be to use domain ontologies in order to establish semantic mappings between different data marts.

6. DESIGN FOR NEW ARCHITECTURES AND APPLICATIONS

Advanced architectures for business intelligence are emerging to support new kinds of applications, possibly involving new and more complex data types. The modeling and design techniques devised so far are mainly targeted towards traditional business applications, and aimed at managing simple alphanumerical data. Thus, it appears inevitable that more general, broader techniques will have to be devised. In this section we discuss the impact of some of the new applications and architectures on modeling and design; other related topics, that we do not address here due to space constraints, are active DWs and DWs for the life sciences.

6.1 Spatial data warehousing

Spatial DWs are characterized by a strong emphasis on spatial data, coming in the form of spatial dimensions or spatial measures. Several works, like [67, 57], show the advantages of using Geographic Information Systems (GIS) characteristics in the analysis of multidimensional data in specific domains. Other works, like [53, 85], implemented more general systems mixing GIS and OLAP.

While all existing conceptual models support basic modeling of a spatial dimension (e.g., most business DWs include a geographic hierarchy built on customers), location data are usually represented in an alphanumeric format. Conversely, picking a more expressive and intuitive representation for these data would reveal patterns that are difficult to discover otherwise.

Preliminary approaches to conceptual modeling for spatial DWs are proposed in [47, 4], where multidimensional models are extended with spatial dimensions, spatial hierarchies, and spatial measures. Also topological relationships and operators such as *intersect* and *inside* as well as user-defined aggregate functions are included to augment the expressivity of these models. From the point of view of logical modeling, the main issue raised by spatial warehousing is how to seamlessly integrate the classical ROLAP and MOLAP solutions (e.g., the star schema) with the specialized data structures used in GISs while preserving high-level performance. In this line, [59] investigates the definition of mappings between the geographical dimension of an OLAP tool and a GIS. Finally, as concerns design methods, adequate solutions for properly moving from conceptual to logical schemata in presence of spatial information must be devised.

6.2 Web warehousing

Web warehouses are DWs that collect Web data. The characteristics of the Web raise new difficulties, mainly due to the semi-structured nature of data, to the lack of control over the sources, and to the frequency of changes on them.

The main challenges in this field are how to integrate heterogeneous web sources and how to automate the process of conceptual design when some or most data sources reside on the Web. Some attempts have been made in this direction, mainly aimed at building a conceptual schema from XML data [35, 81]. In other approaches, like [84, 64], the design of the Web warehouse is driven by frequent user queries and by data quality. Importantly, the development of the Semantic Web opens new exciting possibilities since knowledge is represented according to formal ontologies capable of expressing semantic relationships, which will allow more powerful methods for conceptual design and for data integration to be devised.

6.3 Real-time data warehousing and BPM

As DW systems provide an integrated view of an enterprise, they represent an ideal starting point to build a platform for business process monitoring (BPM). However, performing BPM on top of a DW has a deep impact on design and modeling, since BPM requires extended architectures that may include components not present on stan-

dard DW architectures and may be fed by non-standard types of data (such as data streams). In particular, the fact that BPM implies real-time requirements leads to rethinking ETL components, making the ETL design techniques devised so far questionable. In addition, achieving satisfactory performance for continuous monitoring queries will require more sophisticated logical models for storing data cubes. Arising design issues are summarized in [26]:

- *Right-time design.* While strict real-time will not actually be needed for most applications, data processing must take place in so-called right-time, meaning that information must be ready and complete not later than required by the decision-making process. Thus, a relevant problem for the designer is to understand what is the right-time for the specific business domain.
- *KPI and rule design.* BPM architectures typically include dashboards for viewing key performance indicators (KPIs) and inference engines for managing business rules aimed at giving the decision maker an accurate and timely picture of the business. Hence, suitable techniques for modeling and designing KPIs and business rules, capable of establishing a conceptual connection with the related business goals and of coping with quickly changing requirements, will be necessary.
- *Process design.* In BPM a leading role is played by processes. Hence, BPM design also requires to understand business processes and their relationships in order to find out the relevant KPIs and rules, and to determine where the data to compute them can be found.

6.4 Distributed data warehousing

As in distributed databases, in distributed data warehousing a new phase needs to be added to the design method: the one for designing the distribution, from both the architectural and the physical points of view. During architectural design, general decisions will be taken about which distribution paradigm (P2P, federation, grid) better suits the requirements, how to deploy the DW on the infrastructure, which communication protocols to use, etc. For example, [2] makes the case for a P2P infrastructure for warehousing XML resources, whereas [58] reports how DW systems can be deployed on a grid. On the other hand, the physical point of view mainly addresses how to fragment the DW and how to allocate fragments on the different sites in order to maximize local references to data and to take advantage of the intrinsic parallelism arising from distribution, thus optimizing the overall performance. Though some approaches to fragmentation of DWs have been tempted [54, 23], they are mainly aimed at exploiting local parallelism or at designing ad hoc view fragments for a given workload.

Indeed, distribution is particularly useful in contexts where new data marts are often added, typically because of company mergers or acquisitions. In this case, the most relevant issue is related to integration of heterogeneous data marts as already mentioned in Section 5.

7. CONCLUSION

In this paper we have discussed open issues related to modeling and design of DWs. It is apparent that, though these topics have been investigated for about a decade, several important challenges still arise. Furthermore, ad hoc techniques are required for dealing with the emerging applications of data warehousing and with advanced architectures for business intelligence. Besides, the need for real-time data processing raises original issues that were not addressed within traditional periodically-refreshed DWs. Thus, overall, we believe that research on DW modeling and design is far from being dead, partly because more sophisticated techniques are needed for solving known problems, partly because of the new problems raised during the adaptation of DWs to the peculiar requirements of today's business.

Acknowledgment

We would like to thank the anonymous referees for their careful reading and constructive comments, which helped to improve the presentation. In addition, we would like to warmly thank all the friends who participated in the Dagstuhl Seminar for sharing their ideas with us: Alex Buchmann, Karen Davis, Matteo Golfarelli, Joachim Hammer, Matthias Jarke, Manfred Jeusfeld, Mirek Riedewald, Nick Roussopoulos, Markus Schneider, Timos Sellis, Alkis Simitsis, Dimitri Theodoratos, A Min Tjoa, and Panos Vassiliadis. This paper is based upon our section "Design and Modeling" of an unpublished draft co-authored by all Dagstuhl participants. Our work has been partially supported by the Spanish Research Program PRONTIC and FEDER under project TIN2005-05406, by the Valencia Government (Spain)under DADASMECA, and by the Castilla-La Mancha Government (Spain) under DADS.

8. REFERENCES

[1] A. Abelló, J. Samos, and F. Saltor. YAM^2: a multidimensional conceptual model extending UML. *Information Systems*, 31(6):541–567, 2006.

[2] S. Abiteboul, I. Manolescu, and N. Preda. Constructing and querying peer-to-peer warehouses of XML resources. In *Proc. ICDE*, pages 1122–1123, 2005.

[3] B. Bębel, J. Eder, C. Koncilia, T. Morzy, and R. Wrembel. Creation and management of versions in multiversion data warehouse. In *Proc. ACM SAC*, pages 717–723, 2004.

[4] S. Bimonte, A. Tchounikine, and M. Miquel. Towards a spatial multidimensional model. In *Proc. DOLAP*, pages 39–46, 2005.

[5] M. Blaschka, C. Sapia, and G. Höfling. On schema evolution in multidimensional databases. In *Proc. DaWaK*, pages 153–164, 1999.

[6] M. Bouzeghoub, F. Fabret, and M. Matulovic. Modeling data warehouse refreshment process as a workflow application. In *Proc. DMDW*, 1999.

[7] M. Bouzeghoub and Z. Kedad. A quality-based framework for physical data warehouse design. In *Proc. DMDW*, 2000.

[8] L. Cabibbo and R. Torlone. On the integration of autonomous data marts. In *Proc. SSDBM*, pages 223–231, 2004.

[9] C. Calero, M. Piattini, C. Pascual, and M. A. Serrano. Towards data warehouse quality metrics. In *Proc. DMDW*, 2001.

[10] D. Calvanese, G. De Giacomo, M. Lenzerini, D. Nardi, and R. Rosati. Information integration: Conceptual

modeling and reasoning support. In *Proc. CoopIS*, pages 280–291, 1998.

[11] D. Calvanese, L. Dragone, D. Nardi, R. Rosati, and S. M. Trisolini. Enterprise modeling and data warehousing in Telecom Italia. *Information Systems*, 31(1):1–32, 2006.

[12] J. Cavero, M. Piattini, and E. Marcos. MIDEA: A multidimensional data warehouse methodology. In *Proc. ICEIS*, pages 138–144, 2001.

[13] S. Chaudhuri. Data mining and database systems: Where is the intersection? *IEEE Data Engineering Bulletin*, 21(1):4–8, 1998.

[14] B. Chen, L. Chen, Y. Lin, and R. Ramakrishnan. Prediction cubes. In *Proc. VLDB*, pages 982–993, 2005.

[15] X. Chen and I. Petrounias. Mining temporal features in association rules. In *Proc. PKDD*, pages 295–300, 1999.

[16] J. Feng, Q. Fang, and H. Ding. Prefixcube: Prefix-sharing condensed data cube. In *Proc. DOLAP*, pages 38–47, 2004.

[17] E. Fernandez-Medina, J. Trujillo, R. Villaroel, and M. Piattini. Extending UML for designing secure data warehouses. In *Decision Support Systems*, 2006. In press.

[18] E. Franconi and A. Kamble. A data warehouse conceptual data model. In *Proc. SSDBM*, pages 435–436, 2004.

[19] S. Gardner. Building the data warehouse. *Comm. ACM*, 41(9):52–60, 1998.

[20] P. Giorgini, S. Rizzi, and M. Garzetti. Goal-oriented requirement analysis for data warehouse design. In *Proc. DOLAP*, pages 47–56, 2005.

[21] M. Golfarelli, J. Lechtenbörger, S. Rizzi, and G. Vossen. Schema versioning in data warehouses: enabling cross-version querying via schema augmentation. In *Data & Knowledge Engineering*, 2006. In press.

[22] M. Golfarelli, D. Maio, and S. Rizzi. The Dimensional Fact Model: A conceptual model for data warehouses. *Int. Journ. of Coop. Inf. Syst.*, 7(2-3):215–247, 1998.

[23] M. Golfarelli, V. Maniezzo, and S. Rizzi. Materialization of fragmented views in multidimensional databases. *Data & Knowledge Engineering*, 49(3):325–351, 2004.

[24] M. Golfarelli and S. Rizzi. A methodological framework for data warehouse design. In *Proc. DOLAP*, pages 3–9, 1998.

[25] M. Golfarelli and S. Rizzi. WAND: A CASE tool for data warehouse design. In *Proc. ICDE*, pages 7–9, 2001.

[26] M. Golfarelli, S. Rizzi, and I. Cella. Beyond data warehousing: What's next in business intelligence? In *Proc. DOLAP*, pages 1–6, 2004.

[27] M. Golfarelli, S. Rizzi, and E. Saltarelli. Index selection for data warehousing. In *Proc. DMDW*, pages 33–42, 2002.

[28] J. Han. Towards on-line analytical mining in large databases. *ACM SIGMOD Records*, 27(1):97–107, 1998.

[29] J. Horner, I. Y. Song, and P. P. Chen. An analysis of additivity in OLAP systems. In *Proc. DOLAP*, pages 83–91, 2004.

[30] C. A. Hurtado and A. O. Mendelzon. OLAP dimension constraints. In *Proc. ACM PODS*, pages 169–179, 2004.

[31] B. Hüsemann, J. Lechtenbörger, and G. Vossen. Conceptual data warehouse design. In *Proc. DMDW*, pages 3–9, 2000.

[32] S. Jajodia and D. Wijesekera. Securing OLAP data cubes against privacy breaches. In *Proc. IEEE Symp. on Security and Privacy*, pages 161–178, 2004.

[33] M. Jarke, M. A. Jeusfeld, C. Quix, and P. Vassiliadis. Architecture and quality in data warehouses: An extended repository approach. *Information Systems*, 24(3):229–253, 1999.

[34] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, editors. *Fundamentals of Data Warehousing*. Springer-Verlag, 2000.

[35] M. R. Jensen, T. H. Møller, and T. B. Pedersen. Converting XML DTDs to UML diagrams for conceptual data integration. *Data & Knowledge Engineering*, 44(3):323–346, 2003.

[36] M. E. Jones and I.-Y. Song. Dimensional modeling: identifying, classifying & applying patterns. In *Proc. DOLAP*, pages 29–38, 2005.

[37] M. Kaya and R. Alhajj. Fuzzy OLAP association rules mining based novel approach for multiagent cooperative learning. In *Int. Conf. on Industrial & Engin. Appl. of AI & Expert Syst.*, pages 56–65, 2004.

[38] R. Kimball, L. Reeves, M. Ross, and W. Thornthwaite. *The Data Warehouse Lifecycle Toolkit*. John Wiley & Sons, 1998.

[39] R. Kirkgöze, N. Katic, M. Stolda, and A. M. Tjoa. A security concept for OLAP. In *Proc. DEXA*, pages 619–626, 1997.

[40] L. V. S. Lakshmanan, J. Pei, and Y. Zhao. QC-Trees: an efficient summary structure for semantic OLAP. In *Proc. ACM SIGMOD*, pages 64–75, 2003.

[41] D. Laurent, J. Lechtenbörger, G. Vossen, and N. Spyratos. Monotonic complements for independent data warehouses. *VLDB Journal*, 10(4):295–315, 2001.

[42] J. Lechtenbörger and G. Vossen. Multidimensional normal forms for data warehouse design. *Information Systems*, 28(5):415–434, 2003.

[43] W. Lehner, J. Albrecht, and H. Wedekind. Normal forms for multidimensional databases. In *Proc. SSDBM*, pages 63–72, 1998.

[44] M. Levene and G. Loizou. Why is the snowflake schema a good data warehouse design? *Information Systems*, 28(3):225–240, 2003.

[45] S. Luján-Mora and J. Trujillo. A comprehensive method for data warehouse design. In *Proc. DMDW*, 2003.

[46] S. Luján-Mora, J. Trujillo, and I. Song. A UML profile for multidimensional modeling in data warehouses. In *Data & Knowledge Engineering*, 2006. In press.

[47] E. Malinowski and E. Zimányi. Representing spatiality in a conceptual multidimensional model. In *ACM Int. Work. on GIS*, pages 12–22, 2004.

[48] E. Malinowski and E. Zimányi. Hierarchies in a multidimensional model: From conceptual modeling

to logical representation. *Data & Knowledge Engineering*, 2006. In press.

[49] A. Marotta, F. Piedrabuena, and A. Abelló. Managing quality properties in a ROLAP environment. In *Proc. CAiSE*, pages 127–141, 2006.

[50] J. Mazón, J. Trujillo, M. Serrano, and M. Piattini. Applying mda to the development of data warehouses. In *Proc. DOLAP*, pages 57–66, 2005.

[51] J. Mazón, J. Trujillo, M. Serrano, and M. Piattini. Designing data warehouses: From business requirement analysis to multidimensional modeling. In *Proc. Int. Work. on Requirements Engineering for Business Needs and IT Alignment*, 2005.

[52] MDC. Open Information Model, V1.0. http://www.MDCinfo.com, 1999.

[53] P. Miksovský and Z. Kouba. GOLAP - geographical online analytical processing. In *Proc. DEXA*, pages 442–449, 2001.

[54] D. Munneke, K. Wahlstrom, and M. Mohania. Fragmentation of multidimensional databases. In *Proc. Australasian Database Conference*, pages 153–164, 1999.

[55] T. Niemi, J. Nummenmaa, and P. Thanisch. Normalising OLAP cubes for controlling sparsity. *Data & Knowledge Engineering*, 46(3):317–343, 2003.

[56] OMG. Common warehouse metamodel specification. http://www.omg.org/, 2004.

[57] J. Park and C. Hwang. A design and practical use of spatial data warehouse. In *Proc. IEEE Int. Geoscience and Remote Sensing Symp.*, pages 726–729, 2005.

[58] M. Poess and R. Othayoth. Large scale data warehouses on grid: Oracle database 10g and HP ProLiant systems. In *Proc. VLDB*, pages 1055–1066, 2005.

[59] E. Pourabbas. Cooperation with geographic databases. In M. Rafanelli, editor, *Multidimensional databases: Problems and solutions*, pages 393–432. Idea Group, 2003.

[60] N. Prat, J. Akoka, and I. Comyn-Wattiau. A UML-based data warehouse design method. *Decision Support Systems*, 2006. In press.

[61] T. Priebe and G. Pernul. A pragmatic approach to conceptual modeling of OLAP security. In *Proc. ER*, pages 311–324, 2000.

[62] D. Quass, A. Gupta, I. S. Mumick, and J. Widom. Making views self-maintainable for data warehousing. In *Proc. Int. Conf. on Parallel and Distributed Information Systems*, pages 158–169, 1996.

[63] S. Rizzi et al. Towards a logical model for patterns. In *Proc. ER*, pages 77–90, 2003.

[64] L. I. Rusu, J. W. Rahayu, and D. Taniar. A methodology for building XML data warehouses. *Int. Jour. of Data Warehousing and Mining*, 1(2), 2005.

[65] SAP. Multi-dimensional modeling with BW. Technical report, SAP America Inc. and SAP AG, 2000.

[66] C. Sapia, M. Blaschka, G. Höfling, and B. Dinter. Extending the E/R model for the multidimensional paradigm. In *Proc. ER Workshop on Data Warehousing and Data Mining*, pages 105–116, 1998.

[67] M. Scotch and B. Parmano. SOVAT: Spatial OLAP visualization and analysis tool. In *Proc. HICSS*, 2005.

[68] M. Serrano, C. Calero, J. Trujillo, S. Luján-Mora, and M. Piattini. Empirical validation of metrics for conceptual models of data warehouses. In *Proc. CAiSE*, pages 506–520, 2004.

[69] G. Shankaranarayanan and Y. Cai. Supporting data quality management in decision-making. *Decision Support Systems*, 2006. In press.

[70] A. Simitsis. Mapping conceptual to logical models for ETL processes. In *Proc. DOLAP*, pages 67–76, 2005.

[71] A. Simitsis, P. Vassiliadis, and T. K. Sellis. Optimizing ETL processes in data warehouses. In *Proc. ICDE*, pages 564–575, 2005.

[72] Y. Sismanis, A. Deligiannakis, Y. Kotidis, and N. Roussopoulos. Hierarchical dwarfs for the rollup cube. In *Proc. DOLAP*, pages 17–24, 2003.

[73] Y. Sismanis and N. Roussopoulos. The complexity of fully materialized coalesced cubes. In *Proc. VLDB*, pages 540–551, 2004.

[74] D. Theodoratos and T. Sellis. Designing data warehouses. *Data & Knowledge Engineering*, 31(3):279–301, 1999.

[75] J. Trujillo, S. Luján-Mora, and E. Medina. The Gold model case tool: An environment for designing OLAP applications. In *Proc. ICEIS*, pages 699–707, 2002.

[76] A. Vaisman, A. Mendelzon, W. Ruaro, and S. Cymerman. Supporting dimension updates in an OLAP server. In *Proc. CAiSE*, pages 67–82, 2002.

[77] P. Vassiliadis. *Data Warehouse Modeling and Quality Issues*. PhD thesis, NTUA, Athens, 2000.

[78] P. Vassiliadis, A. Simitsis, P. Georgantas, M. Terrovitis, and S. Skiadopoulos. A generic and customizable framework for the design of ETL scenarios. *Information Systems*, 30(7):492–525, 2005.

[79] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos. Conceptual modeling for ETL processes. In *Proc. DOLAP*, pages 14–21, 2002.

[80] T. Vetterli, A. Vaduva, and M. Staudt. Metadata standards for data warehousing: Open Information Model vs. Common Warehouse Metamodel. *ACM SIGMOD Records*, 3(23), 2000.

[81] B. Vrdoljak, M. Banek, and S. Rizzi. Designing web warehouses from XML schemas. In *Proc. DaWaK*, pages 89–98, 2003.

[82] W. Wang, H. Lu, J. Feng, and J. X. Yu. Condensed cube: An efficient approach to reducing data cube size. In *Proc. ICDE*, pages 155–165, 2002.

[83] R. Winter and B. Strauch. A method for demand-driven information requirements analysis in data warehousing projects. In *Proc. HICSS*, pages 1359–1365, 2003.

[84] J. Zhang, T. W. Ling, R. Bruckner, and A. M. Tjoa. Building XML data warehouse based on frequent patterns in user queries. In *Proc. DaWaK*, pages 99–108, 2003.

[85] L. Zhang, Y. Li, F. Rao, X. Yu, Y. Chen, and D. Liu. An approach to enabling spatial OLAP by aggregating on spatial hierarchy. In *Proc. DaWaK*, volume 2737, pages 35–44, 2003.

[86] J. Zubcoff and J. Trujillo. Extending the UML for designing association rule mining models for data warehouses. In *Proc. DaWaK*, pages 11–21, 2005.