Data warehousing with IBM cloud db2 warehouse

1.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) by the Meta Data Coalition (MDC) and the Common Warehouse Metamodel (CWM) by the OMG (see 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.

2.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.

2.1Spatial 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, show the advantages of using Geographic Information Systems (GIS) characteristics in the analysis of multidimensional data in specific domains. Other works, like , implemented more general systems mixing GIS and OLAP.

While all existing conceptual models support basic modeling of a spatial dimension (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 , 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 (the star schema) with the specialized data structures used in GISs while preserving high-level performance. In this line, 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.

2.2Web 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 . In other approaches, like , 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.

2.3Real-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 standard 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.

2.4Distributed 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, makes the case for a P2P infrastructure for warehousing XML resources, whereas 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 , 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.

3.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 CastillaLa Mancha Government (Spain) under DADS.