Warehousing The World – A Few Remaining Challenges

Torben Bach Pedersen
Aalborg University
tbp@cs.aau.dk

ABSTRACT
Data warehouses (DWs) have become very successful in many enterprises, but only for relatively simple and traditional types of data. It is now time to extend the benefits of DWs to a much wider range of data, making it feasible to literally "warehouse the world". To do this, five unique challenges must be addressed: warehousing data about the physical world, integrating structured, semi-structured, and unstructured data in DWs, integrating the past, the present, and the future, warehousing imperfect data, and ensuring privacy in DWs

Categories and Subject Descriptors
H.2.7 [**Database Management**]: Database Administration – *Data Warehouse and Repository*

General Terms
Algorithms, Performance, Design, Languages, Security

Keywords
Data warehousing, complex data types, models of data, privacy

1. INTRODUCTION
Data warehouses (DWs) have become very successful in many enterprises, by allowing the storage and analysis of large amounts of structured business data. DWs are mostly based on a so-called "multidimensional" data model, where important business events, e.g., sales, are modeled as so-called facts, characterized by a number of hierarchical dimensions, e.g., time and products, with associated numerical measures, e.g., sales price. The multidimensional model is unique in providing a framework that is both intuitive and efficient, allowing data to be viewed and analyzed at the desired level of detail with excellent performance. Traditional data warehouses have worked very well for traditional, so-called *structured* data, but recently enterprises have become aware that DWs are in fact only solving a small part of their real integration and analysis needs.

There is a multitude of different types of data found in most enterprises even today, including structured, relational data, multidimensional data in DWs, text data in documents, emails, and web pages, and semi-structured/XML data such as electronic catalogs. With the current developments within mobile, pervasive and ubiquitous computing, most enterprises will also have to manage large quantities of geo-related data, as well as data from a large amount of sensors. Finally, many analytical models of data have been developed through data mining. The problem with current technologies is that all these types of data/models cannot be integrated and analyzed in a coherent fashion. Instead, applications must develop ad-hoc solutions for integration and analysis, typically for each pair of data types, e.g., relational and text. This obviously is both expensive and error-prone. Privacy protection is, although important, often ignored or given low priority, given the problems with doing the integration and analysis in the first place.

The vision is to develop a breakthrough set of technologies that extend the benefits of DWs to a much wider range of data, making it feasible to literally "warehouse the world". To do this, five unique challenges must be addressed.

The challenges are:

1) Warehousing data about the physical world

2) Integrating structured, semi-structured, and unstructured data in DWs

3) Integrating the past, the present, and the future

4) Warehousing imperfect data

5) Ensuring privacy in DWs

The common base for addressing these challenges could be a new kind of data model, inspired by multidimensional and semi-structured data models, but capable of supporting a much wider range of data. Specifically, support will be added for handling geo-related data (geo models, etc), sensor data (high speed data streams, missing or incorrect values, etc), semi-structured and unstructured data (enabling analysis across structured, semi-structured, and unstructured data), and imperfect (imprecise, uncertain, etc.) data. Support for privacy management will also be built into the framework. In this context, the reseach can explore query languages, query processing/optimization techniques, data integration techniques, and techniques for integrating databases, sensors, and analytical/predictive models of data [1]. Ideally, the contributions would all be integrated into a common prototype system, so that the solutions can be evaluated experimentally using large volumes of real-world data.

This will enable the creation of a *World Warehouse* that provides the same benefits to *all* the described data types as is currently available in traditional DWs for *structured data only*.

The World Warehouse enables the integration and analysis of all types of data using the developed data model and query language. As a distinguishing feature the World Warehouse is protected by an all-encompassing "shield" that provides *integrated privacy management.* All queries to the DW must pass through, and be approved by, the shield, thus ensuring that privacy is not violated.

A large amount of previous work has already been done within the areas of the five challenges, by a large number of researchers.

DOLAP'07, November 9, 2007, Lisbon, Portugal.
ACM 978-1-59593-827-5/07/0011.

Work has been done on aspects of geo warehousing [4,11,12,13] and data streams/sensor data [15], and integrating "pairs" of data types such as (relational, semi-structured), (multidimensional, semi-structured) [5,6,14], (multi-dimensional, text) [9,10], etc. Privacy management is currently a hot topic [2,3], but the special issues related to data warehousing have not been considered in depth yet. Overall, the contributions have not addressed the main issue of integrating and analyzing such diverse types of data coherently and efficiently. It is novel to look at these challenges in combination. Other novel challenges are as follows.

First, the traditional distinction between "real" data values and functions or models that describe data should be broken down. Instead, these two aspects will be seen as a *duality* of the same thing, much like the duality of particles and waves in nuclear physics. The conversion between the two aspects is achieved by *folding* data into models and *unfolding* models into data. The unfolding mechanism means that models/functions can be used in queries just as "real" data values. This unified view will enable much easier integration of past data in databases, present data from sensors, and predicted future data from models.

Second, all data values have an attached uncertainty and imprecision, no matter whether they are "real" historical data or "fake", future, predicted data. Always having a notion of the "imperfection" of the data [7,8,12] also makes it much more natural to compress/aggregate data into patterns/models, e.g., wavelets or Bayesian networks, which can then be unfolded to re-provide the original data.

Third, the idea of folding/unfolding can aid in privacy protection. Privacy can be protected by folding (aggregating/compressing/…) actual data values into *patterns* which is just one kind of function/model describing the data. This of course comes at the cost of some imprecision, but this is also captured natively in the framework. Current approaches to privacy protection such as generalization, condensation, randomization, cloaking, etc., are all special cases of this mechanism, and it is expected that the benefits of a more general approach can be significant.

Fourth, the integrated privacy management "shield" can be enforced by a mechanism based on certification. The idea is that the privacy requirements for a particular data item are built into the data item itself using a special *privacy dimension.* Any query accessing the data item (typically using some kind of aggregation function) will then have to provide a *certificate* that states how the query preserves privacy. The certificate will then be matched against the privacy requirements. If the requirements are met, the data item releases the desired value, otherwise it will refuse to release the value or provide a properly *anonymized* value instead.

2. BRIEF BIO

Torben Bach Pedersen is an Associate Professor at the Department of Computer Science at Aalborg University, Denmark. His research interests include OLAP, multidimensional databases, data integration, location–based services, analysis of web–related data, privacy, data mining and business intelligence applications. He received his Ph.D. in Computer Science from Aalborg University, and his M.S. in Computer Science and Mathematics from Aarhus University.

3. REFERENCES

[1] D. Gawlick. Querying The Past, The Present, and the Future. In *Proc. Of ICDE*, 2004.

[2] J. Gehrke. Models and Methods for Privacy-Preserving Data Publishing and Analysis. In *Proc. Of SIGKDD*, 2006.

[3] G. Gidofalvi, X. Huang, and T. B. Pedersen. Privacy-Preserving Data Mining on Moving Object Trajectories. In *Proc. of MDM*, 2007

[4] C. S. Jensen, A. Kligys, T. B. Pedersen, and I. Timko. Multidimensional Data Modeling For Location-Based Services. *The VLDB Journal*, 13(1):1-21, 2004.

[5] D. Pedersen, J. Pedersen, and T. B. Pedersen. Integrating XML Data in the TARGIT OLAP System. In *Proc. of ICDE*, pp. 778-781, 2004

[6] D. Pedersen, K. Riis, and T. B. Pedersen. XML-Extended OLAP Querying. In *Proc. of SSDBM*, pp. 195--206, 2002.

[7] T. B. Pedersen, C. S. Jensen, and C. E. Dyreson. Supporting Imprecision in Multidimensional Databases Using Granularities. In *Proc. of SSDBM*, pp.90-101, 1999.

[8] T. B. Pedersen. Aspects of Data Modeling and Query Processing for Complex Multidimensional Data. Ph.D. thesis, Faculty of Engineering and Science, Aalborg University, 2000.

[9] J. M. Perez, R. Berlanga, M. J. Aramburu, and T. B. Pedersen. R-Cubes: OLAP Cubes Contextualized With Documents. In *Proc. of ICDE*, 2007.

[10] J. M. Perez, R. Berlanga, M. J. Aramburu, and T. B. Pedersen. Contextualizing Data Warehouses with Documents. To appear in *Decision Support Systems* - Special Issue: Best Papers of DOLAP'05.

[11] I. Timko and T. B. Pedersen. Capturing Complex Multidimensional Data in Location-Based DWs. In *Proc. of ACM-GIS*, pp.147-156, 2004.

[12] I. Timko, C. E. Dyreson, and T. B. Pedersen. Probabilistic Data Modeling and Querying for Location-Based Data Warehouses. In *Proc. of SSDBM*, pp. 273-282, 2005.

[13] I. Timko, C. E. Dyreson, and T. B. Pedersen. Pre-Aggregation with Probability Distributions. In *Proc. of DOLAP*, pp. 35-42, 2006.

[14] X. Yin and T. B. Pedersen. Evaluating XML-Extended OLAP Queries Based on a Physical Algebra. *Journal of Database Management* 17(2):84-114, Special Issue - Best Papers of DOLAP'04, April-June 2006.

[15] X. Yin and T. B. Pedersen. What Can Hierarchies Do for Data Streams? In Proc. of BIRTE, 2007.