Warehousing The World: Challenges From New Types of Data

Torben Bach Pedersen

Department of Computer Science, Aalborg University, Denmark
tbp@cs.aau.dk
http://www.cs.aau.dk/~ tbp

Keynote Abstract

Data warehouses (DWs) have become widely used and successful in many enterprises, by allowing the storage and analysis of large amounts of structured business data. DWs are based on a multidimensional data model, where important business events, e.g., sales, are modeled as 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.

Already today, many different types of data are found in most enterprises, 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. Based on 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 main problem with current technologies is that all these different types of data/models cannot be integrated and analyzed in a coherent fashion. Instead, applications must develop separate 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. Additionally, privacy protection is often given low priority. This situation inspires the vision of developing 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:

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 is envisioned to 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 research can explore query languages, query processing and optimization techniques, data integration techniques, and techniques for integrating databases, sensors, and analytical/predictive models of 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.

Although related work has covered parts of these challenges in isolation, current 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, and instead be seen as a duality of the same thing, much like the duality of particles and waves in nuclear physics. Second, all data values should have a built-in nototion of uncertainty and imprecision, i.e. both "real" historical data and "fake", future, predicted 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. 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. More details about the challenges and the vision can be found in the paper (1)

References

[1] Torben Bach Pedersen. Warehousing The World: A Vision For Data Warehouse Research. In *Kozielski S., Wrembel R. (Eds.): New Trends in Data Warehousing and Data Analysis. Annals of Information Systems*, Vol. 3, 2009.