Humanities Data Warehousing

Janet Delve
University of Portsmouth, UK

INTRODUCTION

Data Warehousing is now a well-established part of the business and scientific worlds. However, up until recently, data warehouses were restricted to modeling essentially numerical data – examples being sales figures in the business arena (in say Wal-Mart's data warehouse (Westerman, 2000)) and astronomical data (for example SKICAT) in scientific research, with textual data providing a descriptive rather than a central analytic role. The lack of ability of data warehouses to cope with mainly non-numeric data is particularly problematic for humanities[1] research utilizing material such as memoirs and trade directories. Recent innovations have opened up possibilities for 'non-numeric' data warehouses, making them widely accessible to humanities research for the first time. Due to its irregular and complex nature, humanities research data is often difficult to model, and manipulating time shifts in a relational database is problematic as is fitting such data into a normalized data model. History and linguistics are exemplars of areas where relational databases are cumbersome and which would benefit from the greater freedom afforded by data warehouse dimensional modeling.

BACKGROUND

Hudson (2001, p. 240) declared relational databases to be the predominant software used in recent, historical research involving computing. Historical databases have been created using different types of data from diverse countries and time periods. Some databases are modest and independent, others part of a larger conglomerate like the North Atlantic Population Project (NAPP) project that entails integrating international census data. One issue that is essential to good database creation is data modeling; which has been contentiously debated recently in historical circles.

When reviewing relational modeling in historical research, (Bradley, 1994) contrasted 'straightforward' business data with incomplete, irregular, complex- or semi-structured historical data. He noted that the relational model worked well for simply-structured business data, but could be tortuous to use for historical data. (Breure, 1995) pointed out the advantages of inputting data into a model that matches it closely, something that is very hard to achieve with the relational model. (Burt[2] and James, 1996) considered the relative freedom of using source-oriented data modeling (Denley, 1994) as compared to relational modeling with its restrictions due to normalization (which splits data into many separate tables), and highlighted the possibilities of data warehouses. Normalization is not the only hurdle historians encounter when using the relational model.

Date and time fields provide particular difficulties: historical dating systems encompass a number of different calendars, including the Western, Islamic, Revolutionary and Byzantine. Historical data may refer to 'the first Sunday after Michaelmas', requiring calculation before a date may be entered into a database. Unfortunately, some databases and spreadsheets cannot handle dates falling outside the late 20th century. Similarly, for researchers in historical geography, it might be necessary to calculate dates based on the local introduction of the Gregorian calendar, for example. These difficulties can be time-consuming and arduous for researchers. Awkward and irregular data with abstruse dating systems thus do not fit easily into a relational model that does not lend itself to hierarchical data. Many of these problems also occur in linguistics computing.

Linguistics is a data-rich field, with multifarious forms for words, multitudinous rules for coding sounds, words and phrases, and also numerous other parameters - geography, educational and social status. Databases are used for housing many types of linguistic data from a variety of research domains - phonetics, phonology, morphology, syntax, lexicography, computer-assisted learning (CAL), historical linguistics and dialectology. Data integrity and consistency are of utmost importance in this field. Relational DataBase Management Systems (RDBMSs) are able to provide this, together with powerful and flexible search facilities (Nerbonne, 1998, introduction).

Bliss and Ritter (2001 IRCS conference proceedings) discussed the constraints imposed on them when using 'the rigid coding structure of the database' developed to house pronoun systems from 109 languages. They observed that coding introduced interpretation of data and concluded that designing 'a typological database is not unlike trying to fit a square object into a round hole. Linguistic data is highly variable, database structures are highly rigid, and the two do not always "fit".' Brown (2001 IRCS conference proceedings) outlined the fact that different database structures may reflect a particular linguistic theory, and also mentioned the trade-off between quality and quantity in terms of coverage.

The choice of data model thus has a profound effect on the problems that can be tackled and the data that can be interrogated. For both historical and linguistic research, relational data modeling using normalization often appears to impose data structures which do not fit naturally with the data and which constrain subsequent analysis. Coping with complicated dating systems can also be very problematic. Surprisingly, similar difficulties have already arisen in the business community, and have been addressed by data warehousing.

MAIN THRUST

Data Warehousing in the Business Context

Data warehouses came into being as a response to the problems caused by large, centralized databases which users found unwieldy to query. Instead, they extracted portions of the databases which they could then control, resulting in the 'spider-web' problem where each department produces queries from its own, uncoordinated extract database (Inmon 2002, pp. 6-14). The need was thus recognized for a single, integrated source of clean data to serve the *analytical needs* of a company.

A data warehouse can provide answers to a completely different range of queries than those aimed at a traditional database. Using an estate agency as a typical business, the type of question their local databases should be able to answer might be 'How many three-bedroomed properties are there in the Botley area up to the value of £150,000?' The type of over-arching question a business analyst (and CEOs) would be interested in might be of the general form 'Which type of property sells for prices above the average selling price for properties in the main cities of Great Britain and how does this correlate to demographic data?' (Begg and Connolly, 2004, p. 1154). To trawl through each local estate agency database and corresponding local county council database, then amalgamate the results into a report would consume vast quantities of time and effort. The data warehouse was created to answer this type of need.

Basic Components of a Data Warehouse

Inmon (2002, p. 31), the 'father of data warehousing', defined a data warehouse as being *subject-oriented, integrated, non-volatile and time-variant.* Emphasis is placed on choosing the right *subjects* to model as opposed to being constrained to model around *applications*. Data warehouses do not replace databases as such - they co-exist alongside them in a symbiotic fashion. Databases are needed both to serve the clerical community who answer day-to-day queries such as 'what is A.R. Smith's current overdraft?' and also to 'feed' a data warehouse. To do this, snapshots of data are extracted from a database on a regular basis (daily, hourly and in the case of some mobile phone companies almost real-time). The data is then transformed (cleansed to ensure consistency) and loaded into a data warehouse. In addition, a data warehouse can cope with diverse data sources, including external data in a variety of formats and summarized data from a database. The myriad types of data of different provenance create an exceedingly rich and varied integrated data source opening up possibilities not available in databases. Thus all the data in a data warehouse is *integrated*. Crucially, data in a warehouse is not updated - it is only added to, thus making it *non-volatile*, which has a profound effect on data modeling, as the main function of normalization is to obviate update anomalies. Finally, a data warehouse has a time horizon (that is contains data over a period) of five to ten years, whereas a database typically holds data that is current for two to three months.

Data Modeling in a Data Warehouse: Dimensional Modeling

There is a fundamental split in the data warehouse community as to whether to construct a data warehouse from scratch, or to build them via data marts. A data mart is essentially a cut-down data warehouse that is