Association for Information Systems
AIS Electronic Library (AISeL)

AMCIS 1999 Proceedings

Americas Conference on Information Systems (AMCIS)

December 1999

Spatial Data in the Data Warehouse: A Nomenclature for Design and Use

Brian Mennecke
Iowa State University

Guy Higgins
The University of Georgia

Spatial Data in the Data Warehouse: A Nomenclature for Design and Use[1]

Brian Mennecke
Iowa State University
mennecke@iastate.edu
Guy Higgins
The University of Georgia
ghiggins@blaze.cba.uga.edu

Introduction

The fields of both computer-based mapping and geographic information systems (GIS) have developed rapidly in the past two decades. However, in many organizations locational data is not easily leveraged by users. The fact that this type of data is collected and stored indicates that it is considered important by the businesses that collect it, but their operational needs often do not entail producing a map or generating geographic queries. Generally, it is a basic operational function -- such as shipping products to the customer -- that necessitates the collection of locational data. Because many organizations have, in effect, out-sourced their basic operational functions to companies like FedEx, they have distanced themselves from many of the impacts of geography. Therefore, in many industry sectors GIS usage is still somewhat limited, if it is used at all.

Efforts by the major GIS venders to encourage the adoption of their software have largely been aimed at applications that use the spatial data held in corporate operational systems; systems that are generally designed for the purpose of providing managers with a means of monitoring and controlling organizational activities. While a number of firms have wholeheartedly adopted and implemented GIS technologies, it is often the case that mapping capabilities are critical to the success of these particular firms' operational systems. In the majority of firms, however, the value added by GIS does not outweigh its costs. The slow pace of GIS adoption in some industry segments is therefore likely due to this misdirected emphasis on positioning GIS technology for use in tasks where its functionality is not critical to achieving organizational success.

Where then does GIS add significant value? Functions such as planning, decision making, tactical and strategic analysis are all areas where GIS functionality can provide the firm with a significant strategic advantage. Most of these functions are more closely associated with decision support systems (DSS) and data warehouses than with operational systems. We believe that while GIS is still a little appreciated technology, it has the potential to be an extremely important technology for decision support.

Yet, to be successful in implementing decision support systems a firm must be able to provide users with data that is standardized, easily accessible, and reliable. For many organizations the data warehouse has become the preeminent tool for managing the data resources that are used in decision support. Although a few venders have made their data warehouse software spatially compatible (e.g., Oracle), many data warehouse environments are not compatible with current GIS software and spatial data. Thus, with regard to spatial data, there is a range of data warehouse implementations that exist in organizations. Because of this, it would be useful to have a nomenclature that could be used to categorize the various implementations that exist in organizations. This paper discusses such as nomenclature as well as the role of GIS within the data warehouse.

Background

Data Warehousing and DSS

Few topics in IS are receiving more discussion today than data warehousing. But what is a data warehouse? According to Inmon and Hackathorn (1994; p. 2), data in a data warehouse are *subject-oriented, integrated, time-variant, and non-volatile.* Thus, a data warehouse is a collection of integrated, summarized, and standardized data optimized for the support of decision-making (Kimball, 1996). The contents of this type of database are organized around specific subject areas (dimensions), frozen at a particular point in time, and, once checked for global consistency and stored, remain unchanged. As the database is optimized for querying rather than data entry, the typical structure of a data warehouse is the dimensional model, or “star join schema.”

In addition to the data warehouse, we suggest that the *data warehouse environment* contains the remaining components of a decision support system (DSS): (1) a model base containing both standardized and ad hoc querying tools, analytical models and tools, and graphical

[1] The review of this paper was managed by Larry West.

display and reporting facilities; (2) the database management system; and (3) the user interface. Thus, the data warehouse environment is an otherwise traditional DSS operating upon the foundation of a data warehouse

The distinction between a data warehouse and a data warehouse environment is made because we believe that it is impossible to separate a discussion of a data warehouse from the decision support system that it supports. Certainly there are important architectural differences between a data warehouse and the traditional operational database. However, just as any discussion of traditional operational databases must begin with the context of that control function in order to make sense of their purpose, a discussion of a data warehouse must acknowledge its ultimate purpose -- decision support.

Spatial Data and GIS

Clearly, we must define what we mean by spatial data and GIS technology before we can explore the incorporation of this technology in a data warehouse environment. Spatial data describe phenomena located on, above, or below the earth's surface and can be considered to be in the "real world" (Laurini & Thompson, 1992). A GIS is a computer-based information system that provides tools to manage, analyze, and display both traditional, or attribute, data and spatial data in an integrated environment. One of the important characteristics of GIS technology is that the software can bring together multiple data sets by "layering" each data set. Layering a group of objects refers to the ability of a GIS to superimpose a number of map features and images. Each layer represents "*a thematic approach to a particular purpose or set of needs,* [and] *may contain one or several different kinds of information*" (Laurini & Thompson, 1992; p. 6). The ability to layer data distinguishes a GIS from other information systems.

The spatial database management tools of a GIS are used to collect and manage both spatially defined data and attribute data. As with any database system, a GIS can query this data and display the results. GIS-based queries can be based on spatial criteria, criteria derived from the attribute data, or some combination of these criteria. Further, the results of these queries can be displayed in both the traditional table format and/or on a map. Spatial queries can include concepts such as "next to," "contained within," and related questions that cannot be executed using other DBMS (Berry, 1993). This provides a powerful set of tools for doing spatial analysis.

The weakness of traditional GIS implementations is that the unique requirements of spatial management systems have not co-existed well with data stored within traditional databases. In fact, GIS are often not compatible with traditional relational database management systems. This is because spatial data are different from traditional data types like numbers or text in that any piece of spatial data has two or more dimensions (e.g., latitude and longitude) while traditional database indexing methods only index one value at a time. As a result of this, GIS venders have had to develop proprietary methods to make their products compatible with other DBMS. While these proprietary methods are functional, they also result in significant data duplication because a separate database, largely translated from the operational database, must be copied and formatted to meet the requirements of the GIS.

GIS offers many advantages to its users. For example, visualization is a powerful functionality of GIS. Complex spatial relationships can much more easily and efficiently be processed by decision makers when they are viewed on a map versus a table (Smelcer & Carmel, 1997). Secondly, there are vast amounts of public and private spatial and attribute data that can be used most effectively only within a GIS environment. In addition, there are many questions that may commonly arise that are almost impossible to answer without a map. Clearly, spatial data and GIS tools are not a necessary part of all DSS; but for many problems the ability to consider data in the context of geography is very helpful. The remainder of this paper considers a nomenclature that can be used to describe the process of moving from an awareness of spatial issues to a full incorporation of GIS technology within a data warehouse.

The Role of Spatial Data in the Data Warehouse

Spatial data can be incorporated into a data warehouse in a multitude of ways and at as much detail as is required by the decision making context. This section presents a nomenclature that can be used to classify how spatial data and spatial analysis can be integrated into the data warehouse environment.

Spatially-Capable Data Warehouse

Just as most operational databases include some spatial data, most data warehouses include some of this same data. We consider these data warehouses to be "*spatially-capable*." A spatially-capable data warehouse is one in which one or more of the dimensions includes locational components. Most data warehouses are spatially-capable because they include dimensions that include information about location (e.g., store address, city, zip codes, etc.). There are no unique or special spatial features present in a spatially-capable data warehouse, there is simply an acknowledgment by the users and designers that there is a spatial dimension to their analytical needs. However, there are no special geographic support features (data elements or data structures) for the spatial elements in the data warehouse or in the DSS. The spatial elements of the spatially-capable data warehouse consist of traditional elements of a relational database systems such as numbers (*i.e.*, a zip code) or text (*i.e*., store name, city name, *etc*.), and these elements can be managed and manipulated within the context of the traditional relational database system (e.g., summarizing records using zip codes). Thus, users would perform spatial analyses within the confines of the DBMS and DSS software that exist in the data warehouse environment. A typical analysis might begin with the extraction of data that include zip codes and conclude with a summarization of one or more attribute fields (e.g., total sales, cost overruns, etc.) based on zip code categories.

Spatially-Enabled Data Warehouse

As the users and designers of a data warehouse recognize a greater need to leverage the spatial dimension of their data, it is reasonable to expect that spatially-focused ad hoc querying tools, spatial analytical models and tools, and spatial display and reporting facilities will be integrated within the data warehouse environment. If we assume that the data warehouse is already spatially capable, then a *spatially-enabled data warehouse* can be created by adding GIS data and tools to the model base of the existing DSS (e.g., by implementing desktop GIS software). In addition, a spatially-enabled data warehouse will require the addition of spatial data (e.g., map data) to the warehouse environment. However, these spatial data will not be fully integrated with the data warehouse, but rather will be managed by the GIS. Since these spatial tables are not fully integrated with the tables of the data warehouse, duplication of data resources will be required.

In summary, a spatially-enabled data warehouse will include not only the data and DSS models that are typically in a data warehouse, but it will also include GIS functionality in the form of separate but compatible GIS tools and data sets that can be used in conjunction with data that are extracted from the data warehouse. A typical analysis scenario in a spatially-enabled data warehouse environment would involve the user first extracting data from the data warehouse using the warehouse's query tools, she would then import these data into the GIS, and finally she would analyze these data in conjunction with the spatial data and tools present in the GIS environment.

Spatial Data Warehouse

In addition to the problem of duplication of data resources due to the lack of a full integration of the tables required by the GIS package and those required by the basic data warehouse function, the requirement for the user to utilize two different sets of querying and analytical tools (e.g., the DBMS query/analysis tools and the spatial query/analysis tools) greatly increases the difficulty of using a spatially-enabled DSS. This is contrary to the purpose of the data warehouse, which is intended to optimize decision support by making the query process very easy for users. A *spatial data warehouse* is a system designed to address these shortcomings by including spatial and aspatial data that are integrated, summarized, and standardized. As in other data warehouse environments, both spatial and aspatial components of the spatial data warehouse are organized around specific subject areas, frozen at a particular point in time, and, once collected and stored, remain unchanged. In addition, this warehouse environment includes spatial and aspatial query tools, analytical models and tools, and display and reporting facilities that support spatial and aspatial analyses.

Because the spatial data are integrated as an integral component of the spatial data warehouse, the spatial data will represent one or more of the dimensions of the data warehouse and, in some cases, one or more fields in the data warehouse's fact table. For example, map data describing the features in a particular region of interest (e.g., the roads, sales territories, and delivery points in a trucking company's data warehouse) would be stored as dimensional data in the spatial data warehouse. These data can then be joined with other attribute data (e.g., items to be delivered to a particular address) in queries to generate maps showing the spatial arrangement of attributes. In addition, however, information about the location of particular resources (e.g., delivery trucks) can be captured (e.g., using GPS receivers) as transactions and included in the data warehouse's fact table. In such a case, the data that would be captured would include the spatial coordinates of the resources (e.g., the location of the truck at a given time period) as well as other transaction data (e.g., the time stamp, the driver ID). Thus, in the spatial data warehouse the spatial data are captured and used in a comparable manner to the way that aspatial data are captured, managed, and used in other data warehouses.

Conclusion

An increasing number of organizations have adopted data warehouse technologies to enable or bolster their decision support capabilities. Because spatial data is an important data resource, many venders and users have begun to leverage spatial data within a data warehouse environment. Yet, no standard nomenclature has been proposed to describe the variety of ways that spatial data can be integrated into a data warehouse. For this reason, we proposed several terms to describe how spatial data could be managed and used in a data warehouse. This nomenclature should be useful for both venders and users to help them to better describe information and technical requirements, vender products, and information services.

References: Available from the authors upon request