ADDING SPATIAL COMPONENTS TO SCIENTIFIC DATA WAREHOUSES

Khaled K. Deeb, Ph.D.
Department of Information Technology, Barry University, 11415 NE 2nd Ave, Miami Shores Florida, 33161

Susan Molina, M.S.
Southeast Fisheries Science Center, National Oceanic and Atmospheric Administration Fisheries
75 Virginia Beach Drive, Miami, Florida 33149

Pam Luckett, Ph.D.
Department of Information Technology, Barry University, 11415 NE 2nd Ave, Miami Shores Florida, 33161

Keywords: Warehouse, Data retrieval, Spatial data components

Abstract: Data warehousing architecture should generally protect the confidentiality of data before it can be published, provide sufficient granularity to enable scientists to variously manipulate data, support robust metadata services, and define standardized spatial components. Data can then be transformed into information that would make them readily available in a common format that is easily accessible, fast, and bridges the islands of dispersed information. The benefits of the warehouse can be further enhanced by adding a spatial component so that the data can be brought to life, overlapping layers of information in a format that is easily grasped by management, enabling them to tease out trends in their areas of expertise.

1 INTRODUCTION

Data are meaningless until they are transformed into information. Over the years many organizations have collected millions of rows of data that, when properly analyzed, translate into useful and profitable information that will assist in deciding the course of future market strategies, the creation of new products, or the retooling of existing production lines to meet consumer demands. In the scientific community, these data are often used to set new policies that can later be made into laws to protect the environment, promote advances in the medical field, or find new natural resources. Since these vast numbers of data can take many hours to process, the databases that house them must meet certain requirements that are often at odds with traditional online transaction processing databases (i.e., OLTP) which require real time access and validation. Hence the birth of the data warehouse, a database that provides mechanisms for synchronizing and updating information obtained from OLTP databases and other external sources, as well as supporting the performance requirements to process and load large volumes of data. The benefits of the warehouse can be further enhanced by adding a spatial component so that the data can be brought to life, overlapping layers of information in a format that is easily grasped by management, enabling them to tease out trends in their areas of expertise.

2 WHAT IS SO SPECIAL ABOUT WAREHOUSE?

Normalized data models evolved to solve the needs of online transactional processing (OLTP) systems that focused on speed of data entry, ease of editing, protection of data integrity by reducing redundancy, and immediate point of entry validation. This model served to solve the problems of the batch processing environment, in

K. Deeb K., Molina S. and Luckett P. (2004).
ADDING SPATIAL COMPONENTS TO SCIENTIFIC DATA WAREHOUSES.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 629-633
DOI: 10.5220/0002668406290633

which data were updated through batches of records that included codes to add, change, or delete from the master flat file. Whenever coding changes were required, the IT group, then called "data management," had to be methodical enough to make sure all occurrences of the offending code where replaced with the new value.

Even though the old flat files were difficult to quality assure, they were very easy for most users to query. Reports generally represented little more than sums across records of data with some additional programming to convert cryptic codes into their more legible legends. Conversely, the OLTP model is perceived as a nightmare for most users who must now write multi table joins across seven or eight tables. Suddenly, the old flat file legacy systems seemed much more desirable. To compound matters, while information technology groups spent their resources designing and deploying elegant data models touting the ease of use of SQL (i.e., Structured Query Language), the user community was left unable to understand the complexities of SQL when confronted with highly normalized designs, and represented to their own management that their data was no longer accessible.

Clearly, there is a need for some compromise between the elegance and efficiency of the OLTP systems, and the needs of the end users to analyze the data that cost so many dollars to capture and ingest. The solution to this dilemma is the data warehouse. Although it raises some brows with apprehension, the data warehouse can be created using the same RDBMS used to house the OLTP database. In fact, a data warehouse is nothing more than a database that has been optimized for retrieval. The main concern of its architects is the delivery of data in a consistent, easy to access format. The emphasis of the design is speed of retrieval.

3 FACTORS THAT MAKE A DATA WAREHOUSE EFFICIENT

As mentioned earlier, a data warehouse must provide good performance, manage large amounts of data, and provide quick ways in which to load large volumes of OLTP data. To support these operations, the manufacturers of data warehousing software have created the following strategies:

Range, Hash, and Composite partitioning: Partitioning involves separating objects into pieces so that they can be managed more easily. Although to the database the object is logically the same, each piece is physically stored separately. Range partitioning is useful when separating data in ranges, such as by year, or by state. On the other hand, when the data cannot be separated into discrete meaningful groups, the data can be separated using a hashing algorithm that ensures that the data will be distributed evenly. The combination of the two partitioning methodologies is called composite partitioning and it uses range partitioning first, while dividing the results using hash partitioning to create sub-partitions (Scherer, 2000, p.145). With any of these techniques, the database designer must analyze both the kinds of data that will be stored in the warehouse, as well as the ways in which the data will be used. Partitioning data based on year may not be effective if the organization typically analyzes data aggregated by state, crossing over multiple years.

Transportable tablespaces: Tablespaces are logical units that hold database objects. Each tablespace may be composed of one or many physical data files. Since data warehouses are used to manage large numbers of data, it may be necessary to move these tablespaces to supply data to a different database or data mart, archive data, or to share data with other databases. Because the tablespaces are based on physical files, there are certain limitations to this capability. These limitations include the requirement that tablespaces be transported only between databases on the same platform, the platform's block size must be the same, and they must use the same character set (Scherer, 2000, p.162).

The star schema: Data warehouses are frequently designed using a dimensional data model called the star schema. In this type of design, there is a central table called a fact table that is related to several look-up tables called dimensions (Silverston, 1997). Fact tables tend to be very large with millions of records and gigabytes to terabytes of data and contain quantitative data. On the other hand, dimension tables tend to be much smaller and contain descriptive data. Both must have a primary key, which for fact tables are usually composites of the foreign keys to the dimension tables. To access tables in this formation, a star query is used, which in turn uses what is referred to as a start join (Scherer, 2000, p. 164). Most RDBMS (i.e., Relational Database Management System) allow developers to use a mechanism called "cost-based optimizer" to allow the computer to optimize the query performance. This is done by regularly compiling statistics on the data and then specifying the "star" hint, which if possible, will cause the database engine to position the largest table, or fact

table, last in the query to reduce the number of rows read with each join (Scherer, 2000 pp. 164-165).

Summary Management and Materialized views: Two important problems faced by data warehousing designers are the aggregation of data across multiple dimensions to hide the complexity of queries, and ensuring that the data are kept up to date when they must be available at all times. Through the use of summary management, performance can be greatly improved by keeping joined and summarized data in objects called materialized views (Scherer, 2000, p.166). In general, a view is a different way in which to look at data. It can be as simple as a single table in which some columns have been renamed or dropped, or as complex as the combination of multiple tables and the creation of summaries. The problem with views is that when they are accessed, the query that they represent must be performed and, depending on its complexity, they can take a long time to execute. On the other hand, materialized views physically store data in this joined and aggregated format, so that when a query is performed, the execution time is minimized to the amount of data stored in the materialized view. With sophisticated RDBMS tools, after statistics have been gathered, the database can recommend which views should be materialized and will even re-write queries if there are materialized views to satisfy a query. The summary management component provides methods to update and propagate the data that make up the materialized view whenever the base tables are updated (Scherer, 2000, p.166).

Before defining the architecture of the geospatial data warehouse it is important for analysts to locate existing stores of data. Especially in large organization, islands of information tend to form and similar data are stored in different formats making the ingestion of data a laborious and time consuming task that detracts from the business of science. Defining the scope of a design effort is critical, since it will lay the ground work for the ultimate design of the warehouse.

4 WHAT IS SO SPATIAL?

The scientific community is becoming increasingly dependent on Geographic Information Systems (GIS). According to J. Michael Fay, GIS experts hold the "key to the future," because by understanding the landscape they can effect positive change on the environment (qtd. In Pratt, 2001). Furthermore, Jack Dangemond, President of ESRI, states that "[t]he application of GIS is limited only by the imagination of those who use it (atd in GIS, 2002). In a nutshell, GIS involves applying layers of information (What is, 2002), whether information about a location's temperature, depth, species of fish caught, number of fishing vessels, any information that can be combined to enable analysts to easily discern patterns in the data. The presentation *Geography Matters* (2002) explains that GIS is not just about maps, but it is the special relationship between data, location information, and the people that manage it. This makes the incorporation of spatial elements into the enterprise data warehouse not just a desirable but a necessary component.

Although GIS has been available for many years, acquiring spatial data layers has not always been simple. Furthermore, the existing GIS analytical tools required levels of expertise that could not be mastered easily by casual users requiring a simple geographical representation of their data. By making spatial shape files available at the enterprise level, scientists will be able to perform comparative studies and visually represent "what-if 'scenarios without expending inordinate amounts of time in the process of ingesting data.

Spatially enabling the warehouse is an ideal solution to provide GIS services to the enterprise. The many standardized layers such as bathymetry and land contours, are generally static and can be loaded one time and remain accessible to all data warehouse customers. This process will result in a great cost savings since currently, since scientists acquiring the data on their own, are loading it into their own databases, and are neither sharing the costs nor the expertise in solving the problem of data ingestion.

Unfortunately, according to Daniel R. Dolk in his paper *Introduction to Modeling Technology and Intelligent Systems Track,* GIS has been largely neglected by IT management and therefore has not benefited from MIS research (Dolk, 1999). Bringing spatial data into the data warehouse represents a unique opportunity for providing benefits to the users by enabling them to share shape files across the enterprise, but they will also benefit from the extensive expertise of their IT department that can leverage existing investments in data acquisition while optimizing the data warehouse.

Large database vendors such as Oracle with its 9i Enterprise RDMS offer a spatial option that enables the storage of spatially enabled data, that is, data that can be interpreted by native SQL and GIS tools. Line geometry can be stored in spatial data object data types that turn the spatial object into a regular field that can be queried using native SQL.

For more complex shape files such as raster data, there is still no support. However, these shape files can still be stored in the data warehouse using binary data types, or, in specialized containers such as Oracle Intermedia, that allow third party products such as E-Spatial to access the data directly from the database. Data stored in binary data format can be easily read by such powerful GIS tools as those produced by the ESRI.

The reason it is important to keep data in a virtual central location (i.e., the warehouse) is that the problem of acquisition is solved a priori by the IT staff in conjunction with the data managers, and last minute research problems the week before presentations at symposia largely go away. By bringing GIS to the warehouse the entire enterprise benefits.

5 DESIGNING THE ETL

ETL or Extract, Transform, and Load procedures are the three steps necessary to populate the data warehouse. Although difficult to set up initially, once the methods are established, the business of synchronizing the transactional database with the warehouse occurs automatically, making the process transparent. Unfortunately, synchronization is slightly more complicated for the scientific community than it is in the business world. The most significant difference is that the sources of scientific data are frequently in the form of data submitted by the owners and is not immediately validated. Depending on the volume of data and the workload of the data management team, it may take from several weeks to several months to make the data ready for the warehouse. Even after the initial constraints are met, specialized analysts can identify outliers that must be corrected in the data warehouse. This presents a problem for ETL, because now the OLTP data needs to be compared with the warehoused data to ensure that the most current version is now available before it is loaded. Therefore, although older data may be archived in the warehouse, it needs to be compared periodically to ensure accuracy. Jaideep Srivastava and Ping-Yao Chen, authors of *Warehouse Creation - A Potential Roadblock,* consider the issue of data loading so important that they have named it a road-block (1999). They have found that generally custom tools must be developed by specialized consultants to handle the ETL issue. It is important to note, however, that once the procedures are put in place, the process of synchronizing the warehouse with the OLTP database occurs automatically.

6 SERVING THE DATA: RETRIEVAL ALTERNATIVES

Using a data warehousing tool that is open and platform independent, affords architects and analysts the capability of opening the data to different tools. Once the data are warehoused, Online Analytical Processing tools can be used to access the data. It is especially important to note that a range of users can be serviced. From the less computer-savvy congressperson who needs broad summarizations to support the passing of legislation, to the expert GIS user who needs granular data to a tenth of a degree, all can benefit from the data stored in the warehouse.

One of the strengths of the data warehouse is its ability to empower end users to perform their own analysis. Historically, users were dependent on the information technology (IT) team to produce inflexible reports that sometimes took months to produce. If an analyst wanted to pivot the report (i.e., turn the report on its side so that the rows in the report become column headings), drill down to the records that contributed to summary data, or simply aggregate data in a different way, another request-for-service document needed to be written, approved, and then submitted to the IT queue until resources became available. On-Line Analytical Processing tools enable users to access multidimensional data on their own, freeing IT resources for other more vital tasks.

Unfortunately, many data warehousing ventures fail because they lose focus on their ultimate goal, which is to serve data to the user community. Joe Kopetsky and Sally Ting of Arcplan Inc. identified three different success factors that will ultimately result in user buy-in: ability, can the stakeholders use the system; willingness, are the users willing to use the system, and knowledge, are the users familiar with how the system works (Kopentsky, 2002). Selecting an OLAP tool that meets these goals and makes the users feel confident that they can get to the data they want, the way the want it, will go a long way to ensuring customer satisfaction.

Serving GIS data poses its own challenges, and these must be met by giving access to the warehoused data at many different levels. There must be simple tools that will enable the entry of natural language queries so that geographic representation of data can be made to management without many hours of labor, while at the same time preserving the expert's ability to use the tools that they have become familiar with over the years.

7 CONCLUSION

Data warehousing solves the problem of data accessibility. Unconcerned with transactional changes, the warehouse architects can concentrate instead on providing a streamlined pipeline to data that can be analyzed expertly. Scientists can benefit from data captured by other scientists without investing the time to locate and reformat the data to make it homogeneous. The warehouse architect does this for them. The traditional relational model, while adequate in managing online transaction processing, falls short when large volumes of data need to be analyzed. To improve accessibility, data warehouses have specific storage and management requirements, such as partitioning, transportable tablespaces, star schema queries and joins, and summary management. Spatial data layers can simply be added to the mix to increase the level of service to the data analysts. As a logical consequence of the need to mine these huge volumes of data, On-Line Analytical Processing tools have been developed to facilitate decision support and obtain business performance metrics. A spatially enabled data warehouse will move cooperative data sharing to a new level by allowing scientists and analysts to manipulate the data for results, rather than as a painful preliminary step to each new study.

REFERENCES

Dolk, D. (March, 1999), *Introduction to Modeling Technology and Intelligent Systems Track.* Paper Presented at the 32"- Hawaii International Conference on System Sciences. Abstract retrieved March 11, 2003 from http://computer.org/proceedings/hicss/0001/00016/00016001.PDF.

ESRI, *GIS & mapping software.* Retrieved *March* 11, 2003 from http://www.esri.com/

Geography Matters. Geographic Information Systems.. Retrieved March 11, 2003, from http./Iwww.gis.com'whatisgis/whatisgis.pdf.

Kopetsky, J. and Ting, S. (2002), If you build it, will they come? The Data Warehousing Institute. Retrieved March 11,2003, from http:/lwww.dw-institute.comfresearch/display.asp?id=5446

Pratt, M (2001, October-December). Real hero gives keynote address. .4rcUser. 4 (4),

Scherer, D., Gaynor, W., Valentinsen, A., & Cursetjee, X. (2000). Oracle 81 Tips and Techniques (pp. 143-188), Berkeley: Oracle Press/McGraw-Hill.

Silverston, L.,Inmon W.H., & Graziano,K. (1997). *The Data Model Resource Book. (pp.* 269-270). New York: Wiley Computer Publishing.

Srivastava, J. and Chen, P. (1999) *Warehouse Creation- A Potential Roadblock to Data Warehousing,* Abstract retrieved March 11, 2003 from http://computer.org/TKDE/tkl999/kOll8abs.htm, 118-126.

What is GIS? (2002). Retrieved March 11, 2002 from http://www.gis.com/whatisgis/index.html