Int. J. Open Problems Compt. Math., Vol. 7, No. 3, September 2014
ISSN 1998-6262; Copyright © ICSRS Publication, 2014
www.i-csrs.org

Requirements engineering for a user centric spatial data warehouse

Vinay Kumar

Professor, Department of IT, VIPS, GGSIPU,
New Delhi 110 088, India
E-mail: vinay5861@gmail.com

Reema Thareja*

Assistant Professor SPM Delhi University,
Author of Oxford University Press, India
reema_thareja@yahoo.com
*Corresponding Author

Received 23rd June 2014; Accepted August 2014

Abstract

Spatial data warehouses (SDW) are used for executing analytical multidimensional queries along with spatial analysis. With recent boom in technology, access and integration of multiple, distributed, heterogeneous and autonomous information sources storing spatial or non-spatial data has become the need. For this purpose, we have proposed a simple star schema to store spatial and non-spatial information in data warehouse. We have also given a framework of activities for collecting data from information sources and organizing it in a geo-spatial data store for querying data and visualizing results. The paper has also discussed the most basic requirements for handling spatial data that will help to develop a user-centric spatial data warehouse.

Keywords: *spatial data warehouse, requirements, operations, schema, hierarchy.*

1 Introduction

For many years, data warehouses are being used for effective utilization of information for business planning and decision making. They store massive amount of subject-oriented data, integrated from several heterogeneous sources over a long period of time [1]. Data warehouses are specifically designed to support knowledge workers to systematically organize and interpret their data to make strategic decisions. Users

perform online analytical processing (OLAP) to view business data from multiple dimensions where each dimension represents some business perspective, like customers, products, time and store. To enable users to analyze data at different levels of summarization, concept hierarchies are defined on each dimension.

However, trend these days is to also exploit spatial data. In [2], authors have claimed that 80% of the overall information stored in computers is geo-spatial related, either explicitly or implicitly. For example, when opening a new store analyzing geo-spatial information in addition to the non spatial information will be highly beneficial. The spatial data analyzed for this purpose could be distance from the residential area, income of people residing in the neighborhood, traffic volume on the roads connecting it, population density in that area to name a few. Therefore, in the last couple of years, efforts are being made to integrate Geographic In- formation Systems (GIS) and OLAP to form SOLAP (Spatial OLAP) for exploring spatial data [3].

A *Spatial data warehouse* (*SDW*) which extends the concept of a traditional DW has an additional spatial component in dimensions or in fact tables and has full support for performing SOLAP operations like finding the store having the largest sales for laptops in New Delhi in 2013, finding the geometry of the region in New Delhi where smart phones usage exceeded that of laptops in the last quarter. For this, the SDWs incorporate *spatial data hierarchies*, *spatial data dimensions* and *spatial measures* within the data warehouse to support *spatial aggregation operations* on them.

In this paper we have talked about the existing star schema used for sales in Section 2. In section 3, we have proposed a star schema which is an extension to the existing one. The proposed schema includes spatial data for analysis. Section 4 details the requirements gathering approach for the proposed schema. It includes activities from collecting user's requirements to identifying relevant maps, collecting them and processing them to be included in the warehouse. The section also lists some basic operations that users like to perform on the processed maps. The paper is concluded in section 5.

2 Existing Star Schema

DWs usually use a star schema that has a large fact table at the centre containing the primary information connected with a number of smaller dimension tables each of which contains information about the entries for a particular attribute in the fact table [4]. The dimension table is joined to the fact table using a primary key to foreign key join, but the dimension tables are not joined to each other.

Fig 1 shows a simple star schema in which the sales fact table is connected with four dimension tables of customers, store, time and product. Data warehouse knowledge workers can use this information to analyze the sales according to the product, time, customer or store.

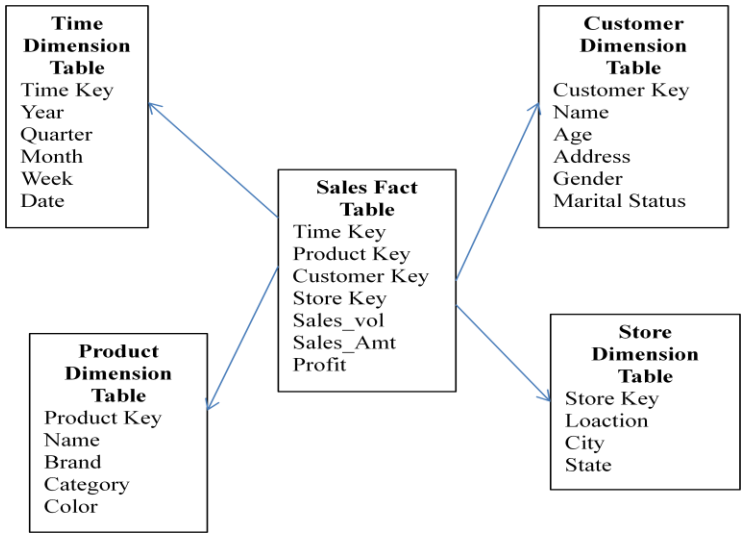

Fig 1: A simple star schema to analyze sales

When a query is issued against the data warehouse schema, the results of the query are produced by joining one or more dimension tables with the fact table. For example, to find the units sales of a product A bought by female customers during the month of May from the stores in New Delhi, a join will be made between the fact table and all the dimension tables.

3 Proposed Star Schema

In this paper, we have proposed a star schema that stores spatial data for analysis. In the previous star schema, the address of the store was stored using non-spatial data– location, city and state. Although, this schema can answer user's queries to a good extent but will fail to answer queries like, what geographical features are restraining from further sales of the product. To answer such questions, we have added some spatial data in the existing schema. The proposed schema is given in fig 2.

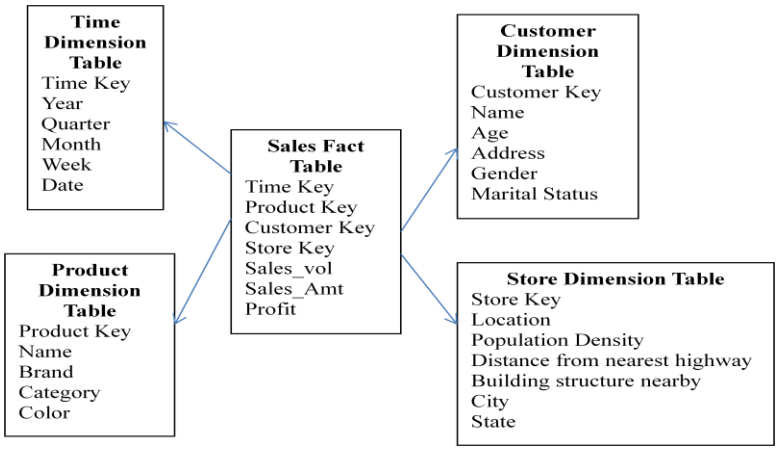

Fig 2: Proposed Star Schema

In the store dimension of the proposed schema, we have added some spatial data like

a. Population density of the area to give an insight into number of people residing in the store's neighborhood
b. Information about the road map that depicts how well connected is the store with major roads of the area
c. Information about land usage to tell the user how the area around store is utilized i.e; existence of large buildings, farmland, etc This gives an insight into the spread of commercial and residential complexes in the nearby area.

4 Requirements Gathering approach for proposed model

In this section, we have designed a requirements engineering framework to understand user's expectations and deliver them spatial data in the requested format. Figure 3 shows the proposed framework for generating spatial information that can be analyzed by users according to the schema shown in fig 2.

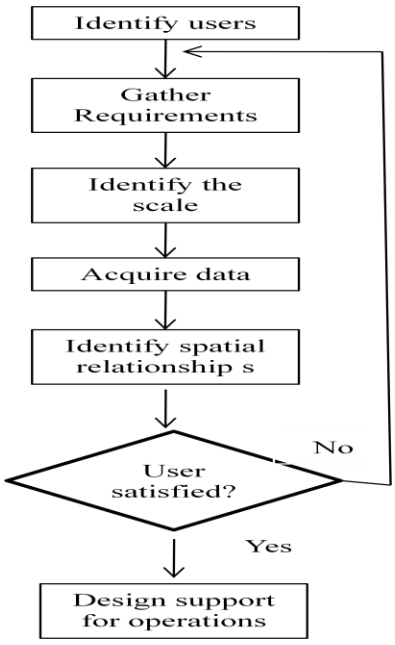

Figure 3: Gathering requirements for the proposed model

4.1 Identify the users

In a data warehouse, there are three types of users who will be working with spatial data. So before gathering requirements, we have first categorized the users under the following types

a. Viewers are users who browse the geographic information occasionally for referential information. They are passive users who expect only ease of use and accessible information.

b. General users use geographical data for decision making process. They are active users so the data warehouse must meet their information requirements.

In DW development team, there has to be GIS specialists (including GIS managers, DBA, programming specialists) for providing support to GIS users.

4.2 Gather requirements

Once the users are identified, their requirements are studied and analyzed. A feasibility study is done to check if their requirements justify the cost and resources of the organization. In case the requirements are within the cost and time budgets, they are documented to lay the foundation of the next step.

4.3 Identify the scale

Attributes pertaining to spatial objects in a map are measured at any of the four levels. So the scale that has to be used to measure the spatial object is specified for each individual spatial object. These levels are

a. Nominal scale is the simplest of all and uses names as labels. For example, nominal scale is used to depict name of highways.
b. Ordinal scale shows order or ranking. For example, such a scale is used to rank areas with respect to their population densities, areas can be ranked according to their commercial use, etc.
c. Interval scale does not have a natural zero but has an arbitrary value. For example, number of stores selling the same product in a given area.
d. Ratio scale unlike the interval scale uses an absolute zero. For example, if number of people residing in an area A is 10000 and in B is 2000, then area A is 5 times more populated than B.

The impact of scales of measurement is reflected on cartographic representation, statistical and spatial data analysis

4.4 Acquire data

The steps for acquiring and making the data useful for analyses purpose is summarized in fig 4.

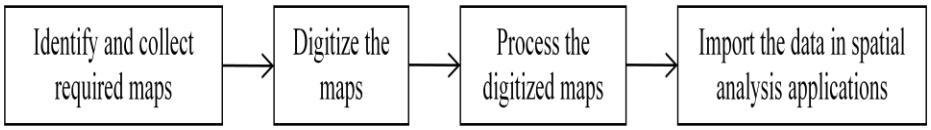

Fig 4: Acquiring spatial data for the data warehouse

Identify and collect required maps: Basically there are two types of maps based on their purpose of use. So before acquiring maps we have first identified which type of map will be required for analysis. The two types of maps are:

a. General purpose or reference maps that are not designed for any specific application. They mainly focus on locations and shows a variety of physical and cultural features like roads, railways, airports, drainage, landforms, forests, build up areas, cultivated areas, etc. these maps are used for geographical referencing like to determine distances, areas and directions.
b. Thematic or special purpose maps are one that depict a particular type of feature or measurement. They are derived from general purpose maps and are produced at different scales. Rivers, highways, population distribution maps are common examples of such maps.

For the sake of spatial data analysis, the reference maps are collected from the mapping agencies working at national, state or provincial level. These agencies prepare such maps either by land surveying or photogrammetric methods. Thematic maps, on the other hand, are usually derived from general purpose maps but they can also be prepared by site investigation, field surveying, remote sensing and other forms of data collection.

Digitize maps: Once the maps are collected, their quality in terms of accuracy, completeness and currency is checked and their control points are identified for registering the output digital data to map coordinates. The next step is digitization of maps so that the coordinates for corresponding objects on the map and on the screen become identical.

Process the digitized map: When the map is scanned for digitization, it captures every point. The resulting raster image is of limited use for analysis so post-scanning data processing is done for vectorization of the raster data. Activities performed during post-scanning data processing include:

a. Conversion of raster image into vector graphics.
b. Conversion of characters in the raster image into alphanumeric data
c. Conversion of cartographic symbols in the raster image into alphanumeric data
d. Graphical data editing for removing data conversion errors
e. Attribute data tagging to add attribute data like feature identifiers, feature codes and counter labels to graphical data in the vector graphics.

Import the data in application: Once the spatial data has been collected, digitized and processed, it is now ready for being incorporated in the application. However, while incorporating the data, coordinate transformation may need to be done if the coordinate system of imported data is different from that of the application. So before integrating the imported data with the existing data, they must be geometrically transformed to the common ground coordinate system.

4.5 Identifying spatial and temporal relationships

Spatial relationships depict the association between different real world objects [11]. Such relationships can be either geometric (adjacent objects share the same boundary) or proximal (one object is close to another). Many analysis applications depend on spatial relationships than on coordinate system. For example, network analysis makes use of spatial relationships between line segments in vector data to determine the

shortest route between two points. Spatial relationships are either stored explicitly (like a line extends "from_node", "to_node" relationships) or computed when required (like determining the point of intersection of two lines, containment of one polygon inside another or computing the position of a point relative to a polygon).

To record the existence and occurrence of spatial objects, time is also stored in data warehouse. This helps the analyst to measure time for objects from their existence to that instance. This is even more helpful for spatial objects that may change over time in terms of space and content. Spatial changes may include a change in location, size, orientation and form thereby altering the relationships that exist between spatial objects.

Once data acquisition phase is complete, data is shown to users, if they are satisfied, the data warehouse development team then moves to the next stage of designing operations otherwise, the process is repeated until the collected data is in tune with users requirements.

4.6 Gathering Technical Requirements for SDW

Blaschka [5] and Pedersen [6], [7] have presented a list of requirements for a multidimensional model spatial data warehouse. The next step is therefore, to identify user's requirements to provide support for their operations. Some basic set of requirements that we have shortlisted include:

a. SDW system should be designed as a *multidimensional* data store to support for increasing data dimensionality over time.

b. SDW should present a simple, intuitive and a user-centric view so that it is easy for end users to understand the structure of the data, analyze it, gather results, visualize and export or save them in different applications.

c. The conceptual design of SDW should be completely independent of its implementation. Such an effective data model enables analyst to view the data at an abstract and high-level, without getting bogged down in details of schema or physical implementation considerations like query optimizations, materialized views, column stores and indexing.

d. The structure of data should be kept separate from its actual values so that OLAP operations are applied to manipulate the multidimensional view of data and its contents independently. For example, dropping a dimension in a data cube does leads to reevaluation of measures in its cells but does not affect the view of the data cube itself for the analyst.

e. SDW should add additional information for descriptive attributes like including one or more *keyword* or *tag* fields or *labels* to identify, qualify and correctly represent spatial objects.

f. SDW should support ragged, unbalanced and uneven hierarchies of data for analysis [8].

g. SDW should provide aggregated data for numerical, statistical, geometric and alphanumeric attributes mot even in one data store but also between data in various data marts.

h. SDW should accurately summarize data. It does not summarize non-additive data and avoids double-counting of data. Moreover, association of spatial geometry to measure values is evaluated while performing aggregations across spatial hierarchies.

i. SDW should support drilling across dimensions to enable sharing of dimensions among different data cubes. Besides, drill across, it also supports drill through capability to allow users to query the *base* data cube to access low-level data stored in databases.

j. SDWs should be are designed to handle updates and deletions over time.

k. The SDW should include not only basic data types (int, char, etc.) but also include complex types like spatial data (point, line, region, etc.), temporal data (time interval, instant, etc.) and abstract *user defined types* (UDTs). It must specify the syntax, semantics and operations that can be performed on those UDTs.

l. Like DW supports generalization and specialization hierarchies on spatial objects. SDW should also provide roll-up operations in hierarchies to answer queries like find the trajectory area where sales of product 'X' has been badly hit.

m. Bedard [9] has suggested that spatial hierarchies in a SDW must be combined as a single dimension

n. Finally, SDW should support operations and aggregations on spatial data like finding topological and cardinal direction relations between interacting spatial objects; manipulation of point, line and region; map generalization, map fusion, sum of areas of selected counties, creating minimum convex polygons, etc.

5 Conclusion

Data warehouses are huge data repositories that integrate data from dispersed heterogeneous databases to enable users to explore data and provide information for making strategic decisions. With increasing need for spatial data for analysis, storing, organizing and utilizing spatial data has become an important issue. Spatial Data Warehouse (SDW) provides a unified view of integrated spatial data from heterogeneous spatial databases, stored cleansed, integrated and transformed to facilitate multiple dimensional spatial data analysis

Therefore, in this paper, we have given a simple star schema that organizes spatial data in the dimension table. We have given a framework for gathering user's requirements, collecting and processing the required maps to finally incorporate them in

data warehouse for spatial analysis. We have also identified a set of requirements for providing users with a set of operations for easily exploring large amounts of spatial data stored multiple levels of granularity.

6 Open Problem

Although spatial OLAP, which is an extension to tradition OLAP is highly desirable to provide fast, flexible, and multidimensional ways for spatial data analysis, it poses great challenges for efficient implementation [10]. Some of these issues include:

a. Selecting appropriate data, integrating it by checking the spatial integrity constraints
b. Incorporating geometric primitives in the metadata [13]
c. Including spatio-temporal indexes, partitioning methods, query optimizers, efficient spatio-temporal topological operators and data update mechanisms in SDW.
d. Integrating spatial data warehouses and spatial data mining with the web technology.

The issues identified needs to be resolved. So our future work will be on web spatial mining.

References

[1] Kimball, R. and Ross, M. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*, 2nd ed., Wiley & Sons, New York 2002.
[2] I. Daratech. Daratech: Geographic information systems markets and opportunities. 2000.
[3] Rivest, S., Bedard, Y., Marchand, P.: Toward better suppport for spatial decision making: Dening the characteristics of spatial on-line analytical processing (SOLAP). Geomatica 55 (2001) 539-555
[4] Thareja R., Data Warehousing, Oxford University Press, India 2009.
[5] Blaschka, M., Sapia, C., Höflng, G. and Dinter, B. 'Finding your way through multidimensional data models', in *9th Int. Workshop on Database and Expert Systems Applications,1998*, p.198.
[6] Pedersen, T. and Jensen, C. 'Multidimensional database technology', *Computer*, 2002, Vol. 34, No. 12, pp.40–46.
[7] Pedersen, T., Jensen, C. and Dyreson, C. 'A foundation for capturing and querying complex multidimensional data', *Information Systems*, 2001, Vol. 26, No. 5, pp.383–423.
[8] Niemi, T., Nummenmaa, J. and Thanisch, P. (2001) 'Logical multidimensional database design for ragged and unbalanced aggregation hierarchies', in *Int. Workshop on Design and Management of Data Warehouses. Interlaken*, Switzerland, Citeseer.
[9] Bédard, Y., Merrett, T. and Han, J. Fundamentals of spatial data warehousing for geographic knowledge discovery', *Geographic Data Mining and Knowledge Discovery*, 2001, Vol. 2, p.53.
[10] Berson, A and Smith, S. J. *Data Warehousing, Data Mining &* OLAP. McGraw-Hill 1997. 612.

[11] Rawlings, J. and Kucera, H. 'Trials and tribulations of implementing a spatial data warehouse', in *Proceedings of the 11th Annual Svmposium on Geographical Information Systems,* Vancouver 1997 510-513.

[12] Chrisman, N. (1997) *Exploring Geographic Information Systems.* John Wiley & Sons, 1997, 320.

[13] Bedard, Y., Gosselin, P., Rivest, S., Proulx, M., Nadeau, M., Lebel, G., & Gagnon, M. (2003). Integrating GIS components with knowledge discovery technology for environmental health decision support. *International Journal of Medical Informatics*, *70*, 79-94.

[14] Fidalgo, R. N., Times, V. C., Silva, J., & Souza, F. (2004). GeoDWFrame: A framework for guiding the design of geographical dimensional schemas. In *LNCS 3181: Proceedings of the 6th International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2004* (pp. 26-37).