Design and Implementation of Enterprise Spatial Data Warehouse

Yin Liang[1, 2] and Hong Zhang[1]

[1]School of Environment Science and Spatial Informatics, China University of Mining and Technology, XuZhou 221008, P.R. China hongzh@cumt.edu.cn
[2]Department of Computer Science and Technology, Xuzhou Normal University, XuZhou 221116, P.R. China liangyinq86@xznu.edu.cn

Abstract. Traditional enterprise data warehouse, an integral part of decision support systems (DSS), provides a variety of granularity data to satisfy requirements of different users. It is estimated that about 80% of the information stored in data warehouse is geo-spatial related. However, traditional data warehouse cannot efficiently process spatial data. With the increasing amount of spatial data stored in spatial databases, how to utilize these spatial data is becoming a critical issue of data warehouse. In this paper, we focus on designing and implementing the enterprise spatial data warehouse for spatial decision-making. We propose three methods of building enterprise spatial data warehouse, and extend traditional enterprise data warehouse model into spatial multidimensional data model, which consists of both spatial and non-spatial dimensions and/or measures. Spatial index with the pre-aggregated results is built on spatial dimension and use the groupings of the index to define a hierarchy. Methods for computation of spatial measure are studied. Extended enterprise spatial data warehouse can accelerate spatial OLAP operations and support the spatial data analysis for decision-making support purposed.

Keywords: *Spatial data warehouse, Spatial measure, Spatial dimension hierarchy*

l. INTRODUCTION

At present, many enterprises have built data warehouse (DW) on which users can carry out their analysis, and obtained much benefit. It is estimated that about 80% of multi-granularity data stored in DW integrates spatial or location information [1], such as supplier address and client address. However, during design and implementation of enterprise data warehouse (EDW), these spatial data is usually represented in an alphanumeric, non-map manner, and lost many spatial characteristic. For example, a certain store address is represented as character string “HuaiHai Street 123”. With the popular use of satellite telemetry system, RS, GPS, and other computerized data collection tools, a huge amount of spatial data has been stored in spatial database (SDB), geographic information systems (GIS), and other spatial information repositories. How to utilize these spatial data to provide analysis environment of more perfect for enterprise decision-making and improve capabilities of spatial data analysis and visualization is becoming a critical issue of DW.

Existing EDWs are neither able to store nor to manipulate spatial data. The management of spatial data is usually carried out by GIS. Therefore, it is an efficient method to combine EDW and GIS to construct enterprise spatial data warehouse (ESDW) for spatial decision-making. On the one hand, DWs Technology can offer efficient access methods and management of a huge amount of data. Furthermore, most on-line analytical processing (OLAP) operations, such as slice, dice, pivot, roll-up, drill-down, and experiences of managing aggregate data, can be used to manage spatial data in SDW. On the other hand, GIS Technology has a long experience in managing spatial data. Especially, spatial index structures, spatial storage management, spatial query and analysis, and spatial information visualization are lucubrated, and applied extensively. However, building ESDW cannot be reduced to simple coupling of EDW and GIS. New techniques for spatial conceptual multidimensional modeling, physical storage, and query optimization are studied to manage high volumes of spatial data. ESDW has been recognized as a key technology for decision-making support [2].

The rest of paper is organized as follows. Section 2 refers to related works. Section 3 details three methods of building ESDW. Section 4 introduces a prototype system. Finally, section 5 gives conclusion and future works of the study.

2. RELATED WORKS

In the ESDW, spatial data and non-spatial data are considered as dimension or measure. In this paper, we present three methods of building ESDW. Before proposing our new methods, we first review related technologies of EDW and GIS.

2.l EDW

In an EDW, data are organized by multidimensional data model. A multidimensional data model is usually represented as a star schema or snowflake schema consisting of a large fact table and a set of smaller dimensional tables joined to the fact table [3]. The fact table stores the primary keys of all the related dimensional tables and numerical measure, such as sales. The dimensional tables can store not only attributes that form a hierarchy, such as day-month-year, but also descriptive attributes, such as store’s name. The dimensions usually represent different analysis perspectives, like customers, product and time, and allow the users to analyze the data from multiple perspectives. Multidimensional data model has been widely applied for non-spatial data, but it is seldom used for spatial data modeling. This is because current multidimensional database technologies do not support the spatial data structures [4]. Therefore, building a multidimensional data model for spatial data is still a challenge. In addition EDW and OLAP tools cannot fully exploit spatial data because spatial data does not have implicit or explicit concept hierarchy.

2.2 GIS

GIS can handle both spatial and non-spatial data to satisfy requirements of some specific domain. GIS not only are powerful tools used to manipulate, manage and visualize spatial databases, but also provide various functions to analyze spatial data. Therefore, the GIS technology is appropriate for a variety of usages including resource management, land surveying, and business planning. However, current GIS cannot effectively and expediently analyze geographical data based on multidimensional data structure. Especially, GIS cannot provide overall information for decision-maker. Since data structure and manipulation of various GIS application are not uniform criterion, resulting in spatial data of the same type in different GIS systems are inconsistent. In this case, it is difficult to obtain overall and consistent data for decision-making support.

2.3 Combining EDW and GIS

The ESDW, which combines EDW and GIS, is built to share spatial information and support decision-making analysis. Recently, SDW is widely studied. In [5], Stefanovic et al. propose a framework of SDW. They extend concept of dimension and measure in DW into spatial dimension and spatial measure. Dimensions in a spatial data cube can be nonspatial dimension, spatial-to-nonspatial dimension and spatial-to-spatial dimension, and measures are both numerical measure and spatial measure. In [6], Rivest et al. extend the definition of spatial measures. In [7], Ferri et al refer to the integration of GIS and DW/OLAP environments. In [8], Fidalgo et al. propose model based on star schema. However, the model does not include the notion of spatial measure, while dimension are classified in a rather complex way. In [1] Bimonte et al. present a multidimensional data model which is able to support complex objects as measures, inter-dependent attributes for measures and aggregation functions. Based on existing model of EDW, spatial information can be integrated into multidimensional data models as dimension or measure to build ESDW.

3. METHODS OF BUILDING ESDW

In this section we present three methods of building ESDW. The first method is the introduction of spatial dimension, the second is introduction of spatial measure, and the third is both spatial dimension and spatial measure.

3.1 Multidimensional Data Model with Spatial Dimension

3.l.l Conceptual Model

A multidimensional model includes spatial dimensions without spatial measures. Spatial dimension can be one or more. If spatial dimension is more than one, a topological relationship is considered. Each spatial dimension includes both description attributes and geometry attributes related to geometry object. Spatial information is represented as spatial dimension, when it is only perspective for analyzing data object property.

When users present a query such as “total sales of products of category A in 2006 in given stores location”, *stores location* is represented as spatial dimension, as shown on Figure 1. In *stores location* table *name*, *city* and *address* are description attributes of store dimension, and location is geometry attribute. This is the case for one spatial dimension.

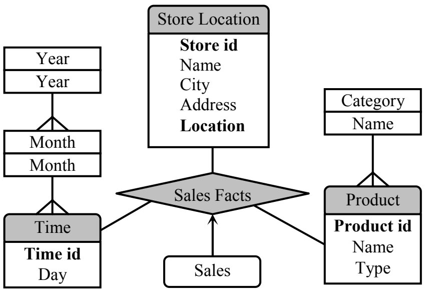

Figure 1. A Star Schema with a Spatial Dimension

If a model includes more than one spatial dimension, a spatial join is required between two or more spatial dimension. Further, the spatial join predicate is specified in the fact table. Figure 2 shows a star schema for the analysis of customer’s buying behavior in every city, which relates two spatial dimensions, *customer* represented as type point and *city location* represented as type area, as well as two non-spatial dimensions, *time* and *product*. Spatial join predicate is *contains* for two spatial dimensions.

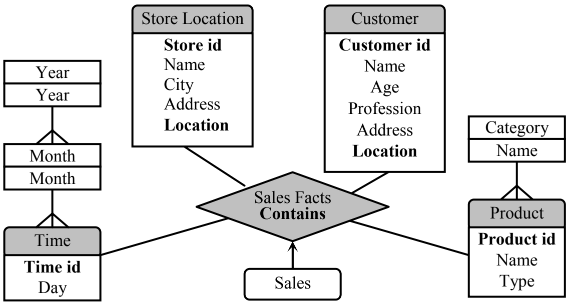

Figure 2. A Star Schema with Two Spatial Dimensions

In this multidimensional model, the aim of multidimensional analysis is sales, location of store, customer and city is different perspectives to analyze sales.

3.l.2 Hierarchy of Spatial Dimension

Each dimension in EDW can be consisted of one or more attribute, and the dimensions are organized as hierarchies of these attributes to represent different degree of generalization, such as day-month-year and city-state-country, etc. In order to improve response performance of OLAP, combinations of different dimensional hierarchies can be pre-computed and stored. Moreover, aggregation results of high dimensional hierarchies can be directly obtained from ones of lower dimensional hierarchies. However, in SDW the spatial dimensions different from non-spatial dimensions since spatial dimensions do not have implicit or explicit concept hierarchies. Set-grouping hierarchies at the spatial dimensions are more complex. During design SDW, dimensional hierarchies may be unknown, or more. Especially, some predefined regions or random ad-hoc query windows created by the users require grouping based on maps, which are computed on the fly. Therefore, we cannot directly apply materialized views techniques widely used in DW to SDW. To solve the problem, two technologies are considered as follows:

1. There may exist some default groupings in some applications. For example, stores in Nanjing are a grouping, and another grouping for stores are covered by Shanghai. Spatial dimensions are organized into multiple hierarchies based on default groupings. If the aggregation results of each default groupings are materialized, the queries that involve these grouping can directly obtain query result.

2. A spatial index is constructed on the objects of the finest granularity on spatial dimension and hierarchy is defined based on the groupings of the index [9]. Each spatial dimension needs to build a spatial index tree.

Figure 3 depicts spatial locations of stores and corresponding R-tree which indexes a set of five store, c1,⋯, c5, whose MBRs are R2 and R3. Based on all the aggregation paths from the bottom to the top of the index tree, data cubes are constructed and conceptual hierarchies on spatial dimension are generated automatically, as shown in Figure 4. By constructing the spatial index tree, we can take advantage of materialized views techniques that exist for EDW to implement spatial views selection, pre-computation, and materialization. This method not only keeps star schema of EDW, but also provides capability to process spatial data.

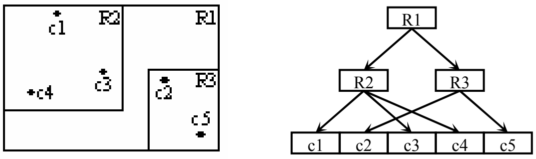

Figure 3. Spatial Data and the Corresponding R-tree

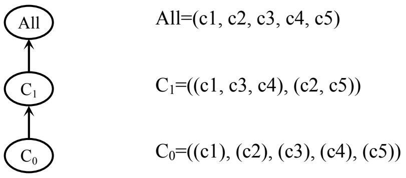
All
All=(c1, c2, c3, c4, c5)
C1
C1=((c1, c3, c4), (c2, c5))
C0
C0=((c1), (c2), (c3), (c4), (c5))

Figure 4. A Spatial Dimensional Hierarchy for the R-tree of Figure 3

In addition, taken spatial index tree as the conceptual hierarchies on spatial dimension, ad-hoc of spatial OLAP can be efficiently processed. In order to obtain aggregation results from non-leaf nodes of spatial index tree and reduce access numbers of nodes to improve query performance, AR-tree, aRB-tree and aCR-tree are used as the hierarchy on spatial dimension.

3.2 Multidimensional Data Model with Spatial Measure

3.2.l Conceptual Model

A multidimensional model includes spatial measures but no spatial dimensions. Spatial measure is usually represented as a collection of spatial pointers to the corresponding spatial objects. Spatial measure is the aim of multidimensional analysis. It is analyzed by non-spatial dimensions. To apply the roll-up and drill-down operations to some dimensional hierarchies, spatial aggregation function should also be defined.

When the user presents the query such as " which cities are customers that buy products of category A in 2006 in?" Geography location of cities is represented as spatial measure. Furthermore, amalgamation aggregation function should be defined to merge border upon cities as a big spatial object. Figure 5 depicts that *City location* is spatial measure and *union* is spatial aggregation function. In this model, *City location* is subject of multidimensional analysis and the users can get information about product sale influenced by geography location of cities.

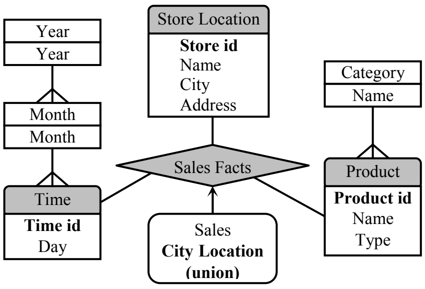

Figure 5. A Star Schema with a Spatial Measure

3.2.2 Computation of Spatial Measure

Spatial measure is similar to numerical measure. According to computing property, aggregation functions for spatial data are also divided into three types of functions [10]: (1) spatial distributive functions, such as convex hull, union, intersection, and length; (2) spatial algebraic functions, such as center of n geometric points or center of gravity; and (3) spatial holistic functions, such as equi-partition and min-distance.

However, spatial measure differs from numerical measure. There are four differences as follows: (1) Numerical measure is simple type and its semantics is limited to quantify description, and spatial measure is complex type; (2) Aggregation result of numerical measure is new numerical value, and aggregation result of spatial measure is a collection of pointers to the corresponding spatial objects and connected spatial objects are merged into a new spatial object; (3) Computing cost of numerical measure is small, and Computing cost of spatial measure is more expensive and it is more required to materialize some spatial queries to improve response time; and (4) Storage space of spatial measure is larger than numerical measure's, and storage space of a spatial object may take kilo- to mega-bytes in storage space. Therefore, it is unpractical to materialize all spatial queries. According to different application, we can use three methods to compute spatial measure as follows:

1. Collection of spatial pointer

Spatial objects are represented by a set of pointers. Advantage of this method is that the storage space is relatively small, and similar to that for non-spatial measures. However, aggregation operations of a group of spatial objects, when necessary, have to be performed on-the-fly. It is a good method if only few spatial objects are aggregated in any pointer collection.

2. Approximate computation of spatial measure

This method is to precompute rough approximation of spatial measure and store. Accuracy of results is not high, but storage space and computing time may be smaller. Due to the users focus on trend change for decision-making analysis, in this case, rough approximation can satisfy requirement of the users. Therefore, this method has been widely studied. The method based on minimum bounding rectangle (MBR) is presented in [5]. The methods based on rotation minimum bounding rectangle (RMBR), multilevel extractive points, and data precision transformation are introduces in [11].

3. Selective materialization of spatial measure

Partial spatial objects selected from objects of aggregation operations are precomputed and stored. Thus, the users can not only obtain accurate results, but also reduce computing time of on-the-fly. Spatial greedy algorithm, pointer intersection algorithm, and object connection algorithm are usually used to determine which sets of spatial objects should be precomputed.

3.3 Multidimensional Data Model with Spatial Dimension and Spatial Measure

The multidimensional data model includes both spatial dimensions and spatial measure. Spatial index tree can be used as set-grouping hierarchy on spatial dimensions. Spatial measure can be not only represented by a collection of pointers to the corresponding spatial objects, but also obtained by applying spatial or topological operators, and is analyzed by both spatial and non-spatial dimensions. When some spatial information is analysis aim, the others are analysis perspectives; this model is a good choice.

Figure 6 illustrates the star schema for the analysis of store location, which closes highway and residential area. The *Highway*, *Store Location*, and *Resident Location* are the spatial dimensions in DW. The fact table is *Sales*, which specify *Distance* operation between spatial dimensions. Aggregation function of spatial measure is *Min-distance*.

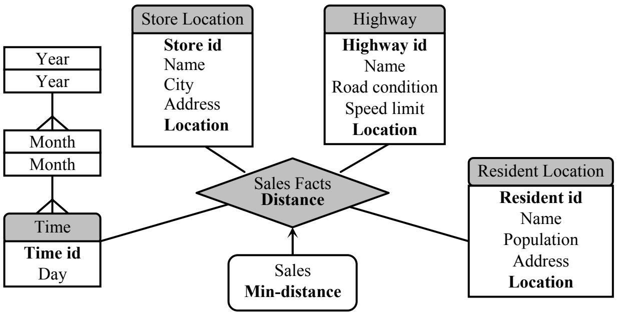

Figure 6. A Star Schema with Spatial Dimension and Spatial Measure

4. A PROTOTYPE

We develop a prototype based on three methods above. The prototype is a three tier architecture. Figure 7 depicts architecture of the prototype. In the prototype, different analytical subject can adopt one of three methods. The designers consider the following three cases to determine which of these models can represent requirements of the users:

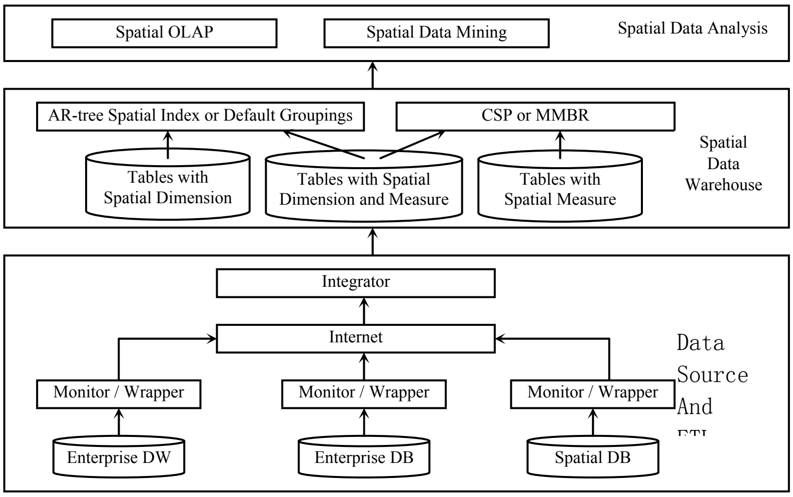

Figure 7. Prototype Architecture

1. If spatial information only needs to be visualized, any one of three models is used to implement. For example, when we find store location with sales more than 500 thousand yuan in 2006, *store location* may be spatial dimension or spatial measure.

2. When the users require comparison of data in different area, or analysis data in some specific area, spatial information is only represented as dimension. The

prototype provides both default groupings and AR-tree as hierarchy on spatial dimensions.

3. If spatial objects need to aggregate location, spatial information is represented as spatial measure, but no spatial dimension. Collection of spatial pointer (CSP) and the method based on minimum bounding rectangle (MMBR) are provided in our prototype.

5. CONCLUSIONS

To integrate spatial information into existing EDW, we propose three methods of building ESDW to improve analysis ability of spatial decision-making. We introduce the corresponding multidimensional data model and key technologies respectively, and apply them to our prototype. The prototype shows feasibility of the methods above. Future study is to further improve query performance for ESDW.

REFERENCES

1. S. Bimonte, A. Tchounikine, and M. Miquel, Towards a spatial multidimensional model, in *Proc. of the 8th ACM international workshop on Data warehouse and OLAP* (2005), pp.39-46.
2. J. Han, R. Altman, V. Kumar, H. Mannila, and D. Pregibon, Emerging scientific applications in data mining, *Communication of the ACM*. Volume 45, Number 8, pp.54-58, (2002).
3. R. Kimball, M. Ross, and R. Merz, *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (John Wiley & Sons: New York, NY, 2002).
4. G. Pestana, M.D. Silva, and Y. Bedard, Spatial OLAP modeling: An overview base on spatial objects changing over time, in *Proc. of ICCC 2005 - IEEE 3rd International Conference on Computational Cybernetics Proceedings* (2005), pp.149-154.
5. N. Stefanovic, J. Han, and K. Koperski, Object-based selective materialization for efficient implementation of spatial data cubes, *IEEE Transactions on Knowledge and Data Engineering*. Volume 12, Number 6, pp.938-958, (2000).
6. S. Rivest, Y. Bedard, and P. Marchand, Toward better support for spatial decision making: Defining the characteristics of spatial on-line analytical processing (SOLAP), *Geomatica*. Volume 55, Number 4, pp.539-555, (2001).
7. F. Ferri, E. Pourabbas, M. Rafanelli, and F. Ricci, Extending geographic databases for a query language to support queries involving statistical data, in *Proc. of the 8th ACM Symposium on Advances in Geographic Information Systems* (2000), pp.220-230.
8. R.N. Fidalgo, V.C. Times, J. Silva, and F. Souza, GeoDWFrame: A framework for guiding the design of geographical dimensional schemas, in *Proc. of the 6th International Conference on Data Warehousing and Knowledge Discovery* (2004), pp.26-37.
9. F.Y. Rao, L. Zhang, X.L. Yu, Y. Li, and Y. Chen, Spatial hierarchy and OLAP-favored search in spatial data warehouse, in *Proc. of the 6th ACM International Workshop on Data Warehousing and OLAP* (2003), pp.48-55.
10. S. Shekhar and S. Chawla, *Spatial Databases: A Tour* (Prentice Hall: New Jersey, 2003).
11. Y.H. Tong, K.Q. Xie, and S.W. Tang, Spatial data warehouse model and spatial data cube computation methods, *Computer Science*. Volume 29, Number 10, pp.1-5, (2002).