Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume II
Data Pro–I

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Typesetter: Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Extending a Conceptual Multidimensional Model for Representing Spatial Data

Elzbieta Malinowski
Universidad de Costa Rica, Costa Rica

Esteban Zimányi
Université Libre de Bruxelles, Belgium

INTRODUCTION

Data warehouses keep large amounts of historical data in order to help users at different management levels to make more effective decisions. Conventional data warehouses are designed based on a multidimensional view of data. They are usually represented as *star* or *snowflake schemas* that contain relational tables called fact and dimension tables. A *fact table* expresses the focus of analysis (e.g., analysis of sales) and contains numeric data called *measures* (e.g., quantity). Measures can be analyzed according to different analysis criteria or *dimensions* (e.g., by product). Dimensions include attributes that can form *hierarchies* (e.g., product-category). Data in a data warehouse can be dynamically manipulated using on-line analysis processing (OLAP) systems. In particular, these systems allow automatic measure aggregations while traversing hierarchies. For example, the roll-up operation transforms detailed measures into aggregated data (e.g., daily into monthly sales) while the drill-down operation does the contrary.

Data warehouses typically include a location dimension, e.g., store or client address. This dimension is usually represented in an alphanumeric format. However, the advantages of using spatial data in the analysis process are well known since visualizing data in space allows users to reveal patterns that are difficult to discover otherwise. Spatial databases have been used for several decades for storing and managing spatial data. This kind of data typically represents geographical objects, i.e., objects located on the Earth's surface (such as mountains, cities) or geographic phenomena (such as temperature, altitude). Due to technological advances, the amount of available spatial data is growing considerably, e.g., satellite images, and location data from remote sensing systems, such as Global Positioning Systems (GPS). Spatial databases are typically used for daily business manipulations, e.g., to find a specific place from the current position given by a GPS. However, spatial databases are not well suited for supporting the decision-making process (Bédard, Rivest, & Proulx, 2007), e.g., to find the best location for a new store. Therefore, the field of spatial data warehouses emerged as a response to the necessity of analyzing high volumes of spatial data.

Since applications including spatial data are usually complex, they should be modeled at a conceptual level taking into account users' requirements and leaving out complex implementation details. The advantages of using conceptual models for database design are well known. In conventional data warehouses, a multidimensional model is commonly used for expressing users' requirements and for facilitating the subsequent implementation; however, in spatial data warehouses this model is seldom used. Further, existing conceptual models for spatial databases are not adequate for multidimensional modeling since they do not include the concepts of dimensions, hierarchies, and measures.

BACKGROUND

Only a few conceptual models for spatial data warehouse applications have been proposed in the literature (Jensen, Klygis, Pedersen, & Timko, 2004; Timko & Pedersen, 2004; Pestana, Mira da Silva, & Bédard, 2005; Ahmed & Miquel, 2005; Bimonte, Tchounikine, & Miquel, 2005). Some of these models include the concepts presented in Malinowski and Zimányi (2004) and Malinowski and Zimányi (2005), to which we will refer in the next section; other models extend non-spatial multidimensional models with different aspects, such as imprecision (Jensen *et al.*, 2004), location-based data (Timko & Pedersen, 2004), or continuous phenomena such as temperature or elevation (Ahmed & Miquel, 2005).

Other authors consider spatial dimensions and spatial measures (Stefanovic, Han, & Koperski, 2000; Rivest, Bédard, & Marchand, 2001; Fidalgo, Times, Silva, & Souza, 2004); however, their models are mainly based on the star and snowflake representations and have some restrictions, as we will see in the next section.

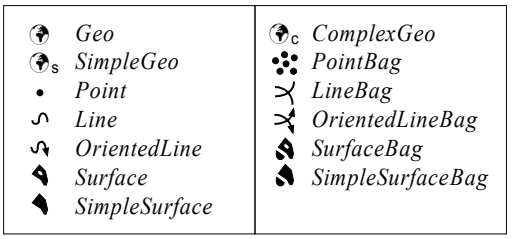
We advocate that it is necessary to have a conceptual multidimensional model that provides organized spatial data warehouse representation (Bédard, Merrett, & Han, 2001) facilitating spatial on-line analytical processing (Shekhar & Chalwa, 2003; Bédard *et al.*, 2007), spatial data mining (Miller & Han, 2001), and spatial statistical analysis. This model should be able to represent multidimensional elements, i.e., dimensions, hierarchies, facts, and measures, but also provide spatial support.

Spatial objects correspond to real-world entities for which the application needs to keep their spatial characteristics. Spatial objects consist of a *thematic* (or descriptive) component and a *spatial* component. The thematic component is represented using traditional DBMS data types, such as integer, string, and date. The spatial component includes its geometry, which can be of type point, line, surface, or a collection of these types. Spatial objects relate to each other with topological relationships. Different topological relationships have been defined (Egenhofer, 1993). They allow, e.g., determining whether two counties touches (i.e., share a common border), whether a highway crosses a county, or whether a city is inside a county.

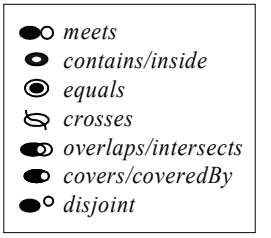
Pictograms are typically used for representing spatial objects and topological relationships in conceptual models. For example, the conceptual spatio-temporal model MADS (Parent *et al.* 2006) uses the pictograms shown in Figure 1.

The inclusion of spatial support in a conceptual multidimensional model should consider different aspects not present in conventional multidimensional models, such as the topological relationships existing between the different elements of the multidimensional model or aggregations of spatial measures, among others. While some of these aspects are briefly mentioned in the literature, e.g., spatial aggregations (Pedersen & Tryfona, 2001), others are neglected, e.g., the influence on aggregation procedures of the topological relationships between spatial objects forming hierarchies.

Figure 1. Pictograms for a) spatial data types and b) topological relationships



(a)



(b)

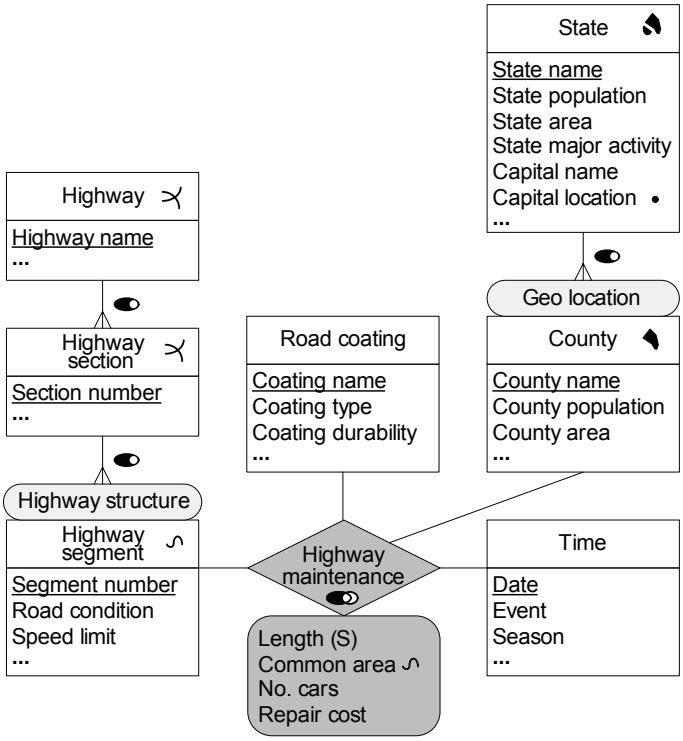
MAIN FOCUS

The MultiDim model (Malinowski & Zimányi, 2008a, 2008b) is a conceptual multidimensional model that allows designers to represent fact relationships, measures, dimensions, and hierarchies. It was extended by the inclusion of spatial support in the different elements of the model (Malinowski & Zimányi, 2004; Malinowski & Zimányi, 2005). We briefly present next our model.

A Conceptual Multidimensional Model for Spatial Data Warehouses

To describe the MultiDim model, we use an example concerning the analysis of highway maintenance costs. Highways are divided into highways sections, which at their turn are divided into highway segments. For each segment, the information about the number of cars and repairing cost during different periods of time is available. Since the maintenance of highway segments is the responsibility of counties through which the highway passes, the analysis should consider the administrative division of the territory, i.e., county and state. The analysis should also help to reveal how the different

Figure 2. An example of a multidimensional schema with spatial elements



types of road coating influence the maintenance costs. The multidimensional schema that represents these requirements is shown in Figure 2. To understand the constructs of the MutiDim model, we ignore for the moment the spatial support, i.e., the symbols for the geometries and the topological relationships.

A *dimension* is an abstract concept for grouping data that shares a common semantic meaning within the domain being modeled. It represents either a level or one or more hierarchies. *Levels* correspond to entity types in the entity-relationship model; they represent a set of instances, called *members*, having common characteristics. For example, Road Coating in Figure 2 is a one-level dimension.

Hierarchies are required for establishing meaningful paths for the roll-up and drill-down operations. A hierarchy contains several related levels, such as the County and State levels in Figure 2. They can express different structures according to an *analysis criterion*, e.g., geographical location. We use the criterion name to differentiate them, such as Geo location, or Highway structure in Figure 2.

Given two related levels of a hierarchy, one of them is called *child* and the other *parent* depending on whether they include more detailed or more general data, respectively. In Figure 2, Highway segment is a child level while Highway section is a parent level. A level of a hierarchy that does not have a child level is called *leaf* (e.g., Highway segment); the level that does not have a parent level is called *root* (e.g., Highway).

The relationships between child and parent levels are characterized by *cardinalities*. They indicate the minimum and the maximum numbers of members in one level that can be related to a member in another level. In Figure 2, the cardinality between the County and State levels is many-to-one indicating that a county can belong to only one state and a state can include many counties. Different cardinalities may exist between levels leading to different types of hierarchies (Malinowski & Zimányi, 2008a, 2008b).

Levels contain one or several *key attributes* (underlined in Figure 2) and may also have other *descriptive* attributes. Key attributes indicate how child members are grouped into parent members for the roll-up operation. For example, in Figure 2 since State name is the

key of the State level, counties will be grouped according to the state name to which they belong.

A *fact relationship* (e.g., Highway maintenance in Figure 2) represents an n-ary relationship between leaf levels. It expresses the focus of analysis and may contain attributes commonly called *measures* (e.g., Repair cost in the figure). They are used to perform quantitative analysis, such as to analyze the repairing cost during different periods of time.

Spatial Elements

The MultiDim model allows including spatial support for levels, attributes, fact relationships, and measures (Malinowski & Zimányi, 2004; Malinowski & Zimányi, 2005).

Spatial levels are levels for which the application needs to keep their spatial characteristics. This is captured by its geometry, which is represented using the pictograms shown in Figure 1 a). The schema in Figure 2 has five spatial levels: County, State, Highway segment, Highway section, and Highway. A level may have spatial attributes independently of the fact that it is spatial or not, e.g., in Figure 2 the spatial level State contains a spatial attribute Capital.

A *spatial dimension* (respectively, *spatial hierarchy*) is a dimension (respectively, a hierarchy) that includes at least one spatial level. Two linked spatial levels in a hierarchy are related through a topological relationship. These are represented using the pictograms of Figure 1 b). By default we suppose the *coveredBy* topological relationship, which indicates that the geometry of a child member is covered by the geometry of a parent member. For example, in Figure 2, the geometry of each county must be covered by the geometry of the corresponding state. However, in real-world situations different topological relationships can exist between spatial levels. It is important to consider these different topological relationships because they determine the complexity of the procedures for measure aggregation in roll-up operations (Malinowski & Zimányi, 2005).

A *spatial fact relationship* relates two or more spatial levels, e.g., in Figure 2 Highway maintenance relates the Highway segment and the County spatial levels. It may require the inclusion of a spatial predicate for spatial join operations. For example, in the figure an intersection topological relationship indicates that users focus their analysis on those highway segments that intersect counties. If this topological relationship is not included, users are interested in any topological relationships that may exist between them.

A (spatial) fact relationship may include measures, which may be spatial or thematic. *Thematic measures* are usual numeric measures as in conventional data warehouses, while *spatial measures* are represented by a geometry. Notice that thematic measures may be calculated using spatial operators, such as distance, area, etc. To indicate that a measure is calculated using spatial operators, we use the symbol (S). The schema in Figure 2 contains two measures. Length is a thematic measure (a number) calculated using spatial operators; it represents the length of the part of a highway segment that belongs to a county. Common area is a spatial measure representing the geometry of the common part. Measures require the specification of the function used for aggregations along the hierarchies. By default we use *sum* for numerical measures and *spatial union* for spatial measures.

Different Approaches for Spatial Data Warehouses

The MultiDim model extends current conceptual models for spatial data warehouses in several ways. It allows representing spatial and non-spatial elements in an orthogonal way; therefore users can choose the representation that better fits their analysis requirements. In particular we allow a non-spatial level (e.g., address represented by an alphanumeric data type) to roll-up to a spatial level (e.g., city represented by a surface). Further, we allow a dimension to be spatial even if it has only one spatial level, e.g., a State dimension that is spatial without any other geographical division. We also classify different kinds of spatial hierarchies existing in real-world situations (Malinowski & Zimányi, 2005) that are currently ignored in research related to spatial data warehouses. With respect to spatial measures, we based our approach on Stefanovic *et al.* (2000) and Rivest *et al.* (2001); however, we clearly separate the conceptual and the implementation aspects. Further, in our model a spatial measure can be related to non-spatial dimensions as can be seen in Figure 3.

For the schema in Figure 3 the user is interested in analyzing locations of accidents taking into account the different insurance categories (full coverage, partial coverage, etc.) and particular client data. The model includes a spatial measure representing the location of an accident. As already said above, a spatial function is

Figure 3. Schema for analysis of accidents with a spatial measure location

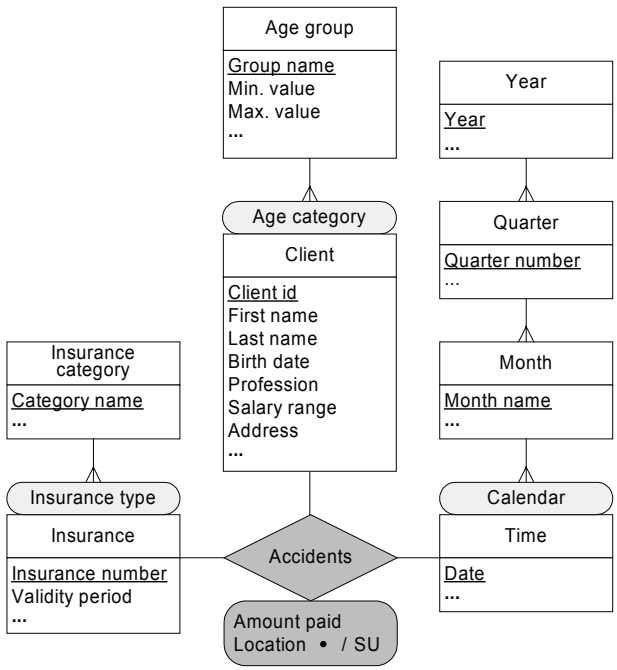
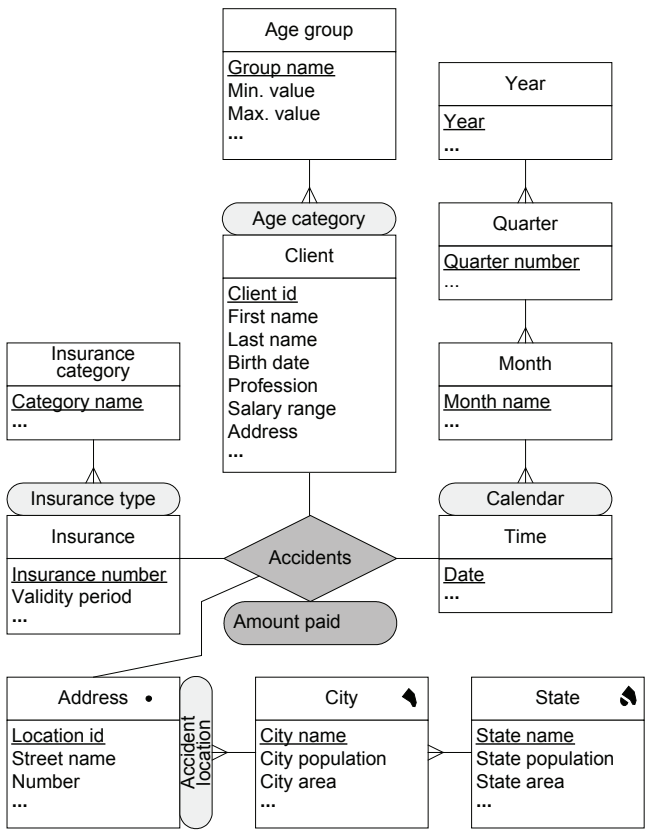


Figure 4. Spatial dimension: another variant of a schema for analysis of accidents



needed to aggregate spatial measures through hierarchies. By default the spatial union is used: when a user rolls-up to the Insurance category level, the locations corresponding to different categories will be aggregated and represented as a set of points. Other spatial operators can be also used, e.g., center of n points.

Other models (e.g., Fidalgo *et al.*, 2004) do not allow spatial measures and convert them into spatial dimensions. However, the resulting model corresponds to different analysis criteria and answers to different queries. For example, Figure 4 shows an alternative schema for the analysis of accidents that results from transforming the spatial measure Location of the Figure 3 into a spatial dimension Address.

For the schema in Figure 4 the focus of analysis has been changed to the amount of insurance paid according to different geographical locations. Therefore, using this schema, users can compare the amount of insurance paid in different geographic zones; however, they cannot aggregate locations (e.g., using spatial union) of accidents as can be done for the schema in Figure 3. As can be seen, although these models are similar, different analyses can be made when a location is handled as a spatial measure or as a spatial hierarchy. It is the designer's decision to determine which of these models better represents users' needs.

To show the feasibility of implementing our spatial multidimensional model, we present in Malinowski and Zimányi, (2007, 2008b) their mappings to the object-relational model and give examples of their implementation in Oracle 10g.

FUTURE TRENDS

Bédard *et al.*, (2007) developed a spatial OLAP tool that includes the roll-up and drill-down operations. However, it is necessary to extend these operations for different types of spatial hierarchies (Malinowski & Zimányi, 2005).

Another interesting research problem is the inclusion of spatial data represented as continuous fields, such as temperature, altitude, or soil cover. Although some solutions already exist (Ahmed & Miquel, 2005), additional research is required in different issues, e.g., spatial hierarchies composed by levels representing field data or spatial measures representing continuous phenomena and their aggregations.

Another issue is to cope with multiple representations of spatial data, i.e., allowing the same real-world object to have different geometries. Multiple representations are common in spatial databases. It is also an important aspect in the context of data warehouses since spatial data may be integrated from different source systems that use different geometries for the same spatial object. An additional difficulty arises when the levels composing a hierarchy can have multiple representations and one of them must be chosen during roll-up and drill-down operations.

CONCLUSION

In this paper, we referred to spatial data warehouses as a combination of conventional data warehouses and spatial databases. We presented different elements of a spatial multidimensional model, such as spatial levels, spatial hierarchies, spatial fact relationships, and spatial measures.

The spatial extension of the conceptual multidimensional model aims at improving the data analysis and design for spatial data warehouse and spatial OLAP applications by integrating spatial components in a multidimensional model. Being platform independent, it helps to establish a communication bridge between users and designers. It reduces the difficulties of modeling spatial applications, since decision-making users do not usually possess the expertise required by the software used for managing spatial data. Further, spatial OLAP tools developers can have a common vision of the different features that comprise a spatial multidimensional model and of the different roles that each element of this model plays. This can help to develop correct and efficient solutions for spatial data manipulations.

REFERENCES

Ahmed, T., & Miquel, M. (2005). Multidimensional structures dedicated to continuous spatio-temporal phenomena. *Proceedings of the 22nd British National Conference on Databases*, pp. 29-40. Lecture Notes in Computer Science, N° 3567. Springer.

Bédard, Y., Merrett, T., & Han, J. (2001). Fundamentals of spatial data warehousing for geographic knowledge

discovery. In J. Han & H. Miller (Eds.), *Geographic Data Mining and Knowledge Discovery,* pp. 53-73. Taylor & Francis.

Bédard, Y., Rivest, S., & Proulx, M. (2007). Spatial On-Line Analytical Processing (SOLAP): Concepts, Architectures and Solutions from a Geomatics Engineering Perspective. In R. Wrembel & Ch. Koncilia (Eds.), *Data Warehouses and OLAP: Concepts, Architectures and Solutions*, pp. 298-319. Idea Group Publishing.

Bimonte, S., Tchounikine, A., & Miquel, M. (2005). Towards a spatial multidimensional model. *Proceedings of the 8th ACM International Workshop on Data Warehousing and OLAP,* 39-46.

Egenhofer, M. (1993). A Model for detailed binary topological relationships, *Geomatica*, *47*(3&4), 261-273.

Fidalgo, R., Times, V., Silva, J., & Souza, F. (2004). GeoDWFrame: A framework for guiding the design of geographical dimensional schemes. *Proceedings of the 6th International Conference on Data Warehousing and Knowledge Discovery,* pp. 26-37. Lecture Notes in Computer Science, N° 3181. Springer.

Jensen, C.S., Klygis, A., Pedersen, T., & Timko, I. (2004). Multidimensional Data Modeling for Location-Based Services. *VLDB Journal, 13*(1), pp. 1-21.

Malinowski, E. & Zimányi, E. (2004). Representing Spatiality in a Conceptual Multidimensional Model. *Proceedings of the 12th ACM Symposium on Advances in Geographic Information Systems,* pp. 12-21.

Malinowski, E. & Zimányi, E. (2005). Spatial Hierarchies and Topological Relationships in the Spatial MultiDimER model. *Proceedings of the 22nd British National Conference on Databases*, pp. 17-28. Lecture Notes in Computer Science, N° 3567. Springer.

Malinowski, E. & Zimányi, E. (2007). Spatial Hierarchies and Topological Relationships in the Spatial MultiDimER model. *GeoInformatica*, 11(4), 431-457.

Malinowski, E. & Zimányi, E. (2008a). Multidimensional Conceptual Models, *in this book*.

Malinowski, E. & Zimányi, E. (2008b). *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications*. Springer.

Pedersen, T.B. & Tryfona, N. (2001). Pre-aggregation in spatial data warehouses. *Proceedings of the 7th International Symposium on Advances in Spatial and Temporal Databases*, pp. 460-480. Lecture Notes in Computer Science, N° 2121. Springer.

Miller, H. & Han, J. (2001) *Geographic Data Mining and Knowledge Discovery.* Taylor & Francis.

Parent, Ch., Spaccapietra, S., & Zimányi, E. (2006). *Conceptual modeling for traditional and spatio-temporal applications: The MADS approach.* Springer.

Pestana, G., Mira da Silva, M., & Bédard, Y. (2005). Spatial OLAP modeling: An overview based on spatial objects changing over time. *Proceedings of the IEEE 3rd International Conference on Computational Cybernetics*, pp. 149-154.

Rivest, S., Bédard, Y., & Marchand, P. (2001). Toward better support for spatial decision making: defining the characteristics of spatial on-line analytical processing (SOLAP). *Geomatica*, *55*(4), 539-555.

Shekhar, S., & Chawla, S. (2003). *Spatial Databases: A Tour*. Prentice Hall.

Stefanovic, N., Han, J., & Koperski, K. (2000). Object-based selective materialization for efficient implementation of spatial data cubes. *Transactions on Knowledge and Data Engineering, 12*(6), pp. 938-958.

Timko, I. & Pedersen, T. (2004). Capturing complex multidimensional data in location-based data warehouses. *Proceedings of the 12th ACM Symposium on Advances in Geographic Information Systems,* pp. 147-156.

KEY TERMS

Multidimensional Model: A model for representing the information requirements of analytical applications. It comprises facts, measures, dimensions, and hierarchies.

Spatial Data Warehouse: A data warehouse that includes spatial dimensions, spatial measures, or both, thus allowing spatial analysis.

Spatial Dimension: An abstract concept for grouping data that shares a common semantics within the

domain being modeled. It contains one or more spatial hierarchies.

Spatial Fact Relationship: An n-ary relationship between two or more spatial levels belonging to different spatial dimensions.

Spatial Hierarchy: One or several related levels where at least one of them is spatial.

Spatial Level: A type defining a set of attributes, one of them being the geometry, keeping track of the spatial extent and location of the instances, or members, of the level.

Spatial Measure: An attribute of a (spatial) fact relationship that can be represented by a geometry or calculated using spatial operators.