Spatial Data Warehousing: A Strategy for Integrated Urban Data Management in Support of Decision Making

C. Vincent Tao

Department of Geomatics Engineering, The University of Calgary
2500 University Dr. NW, Calgary, Alberta, Canada, T2N 1N4

Abstract
Various data technologies for spatial data management under urban environments are briefly reviewed. The concept and principle of spatial data warehousing with respect to the urban data environment is given. The characteristics of the spatial data warehouse and its architecture are described. The potential use of spatial data warehousing for the development of an integrated urban data management in support of decision making is discussed. A three-tiered architecture for building a spatial data warehouse is then proposed. Finally, issues involving the design and implementation of spatial data warehouses are addressed.

I. INTRODUCTION

Computers have been applied in urban planning and management almost since their inception, but only recently with the development of graphics, distributed processing, and network communications has software emerged which can now be used routinely and effectively. Geographic or Geospatial Information Systems (GIS) has been the cornerstone of many urban data strategies. Recently, due to the strong impact of Information Technologies (IT), the use of spatial data warehousing strategy for managing large and heterogeneous spatial databases for urban applications has gained increasing attention. Spatial data warehouse is essentially a collection of large amounts of historical data as well as a collection of decision support tools that can be integrated for decision support.

In this paper, followed by a review of various data technologies for spatial data management under urban environments, a new strategy to integrated data management in support of decision making under the urban environment is proposed. The paper addresses the needs of the development of a new data strategy based on the spatial data warehousing concept. An architecture of an integrated data management for urban environments is presented. Various issues regarding the design and implementation of a spatial data warehouse are discussed.

II. THE EVOLUTION OF URBAN DATA MANAGEMENT

Efficient data management has been one of the primary concerns for many urban applications, such as land registration, housing management, transportation control, and utility management, etc. Data strategies applied to urban data management can be considered to be evolved in three stages:

- *transaction-based data processing*
- *database-based data management and processing*
- *data warehouse-based management and decision making*

Transaction-Based Data Processing

Information systems designed only to automate existing manual procedures are called *transaction-based systems*. This kind of systems takes an input record (transaction) and processes it through a series of programs that manipulate the data on the record, and then write new information into the files (Figure 1). They are essentially file-based processing systems.

The transaction-based systems perform efficiently in most cases due to its focussed design and relatively small amount of data handled. As the expansion of the systems is required or more programs and files are added, the systems become complex and difficult to maintain. Gradually, the transaction-based design loses its advantage in performance, due to the complicated processing flow and poor organization of data files. Consequently, these systems became of a problem to municipal governments. Figure 2 illustrates an example of the complex data processing flow. One can imagine that the system like this would be hardly maintained and updated.

1082-4006/00/0602-113$5.00

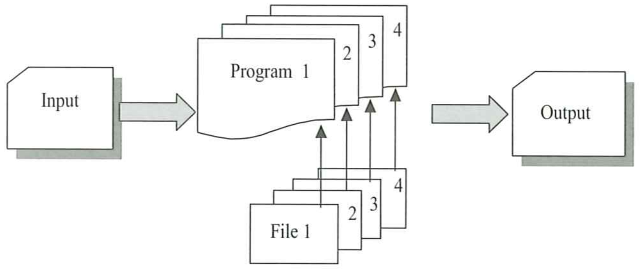

Figure 1. The architecture of a transaction-based system

Database-Based Data Management and Processing

Development of database management systems (DBMS) provides a solution to the above problem. DBMS makes the data independent from the programs, applications and systems. With a database managing urban data, various programs and applications can be built without knowing the physical structures of the data stored in the databases. Having this architecture, data updating or modification does not affect the programs and the changes made to the programs also do not alter the data. It eases the implementation of application programs and greatly reduces the data redundancy. Moreover, it allows the establishment of a centralized data repository to control data quality, integrity and security (Figure 3).

Compared to the transaction-based systems, the development cycle of such systems is much longer. Moreover, implementation cost is of considerable high with a difficulty on cost-justification. This is because that such a system is used by more than one department and is beneficial to many sectors. The cost-savings by these related departments are not easily accountable. This was one of the main reasons on preventing people developing database-based solutions. This is not a problem today due to the dramatically decreased costs of hardware, software and databases.

In most municipalities, the strategy for building a centralized database is adopted to only manage the base mapping data such as topographic data, survey controls, land records, road networks etc., since these data sets are considered as core data that must be shared by most departments. While other spatial and non-spatial data sets are acquired, managed and maintained by the associated departments as part of their business practice. For example, the environment services department is responsible for its own wastes and contaminated data. With this data strategy, core data can be maintained and updated by a specialized data processing unit in order to ensure the integrity, quality and currency of the data. It leaves the flexibility of the use of data and the application development to the departments.

In fact, many departments have designed and implemented their own operation systems for their specific applications such as transportation systems, utility management systems, and emergency response systems etc. Although these systems use the shared base mapping data as input, the data has been extracted, transformed or cleaned in a way that it fits the specific needs of the application. Often, the shared data has been integrated or combined with the department's legacy data, spatial or non-spatial, and finally stored in an operational database. As a result, many operational systems for various urban applications are developed which results in a two-tiered architecture, shown in Figure 4. This architecture is very common in current urban data management environments.

Data Warehouse-Based Data Management and Decision Support

With the further development of urban management and applications, functions such as planning and decision making become more and more important to the sustainable urban development. As stated in Huxhold (1991), urban information system should be designed and implemented in a way that it can provide information for all of the three levels, operations, management and policy.

Decision support is a complex process, for instance, strategic planning, scheduling of operations, and investment appraisal. In order to support the decision making process, the system should be able to use data from various departments and information from many stages of business practices. The system also should provide sophisticated and powerful analytical tools to assist the decision makers for problem solving. Finally,

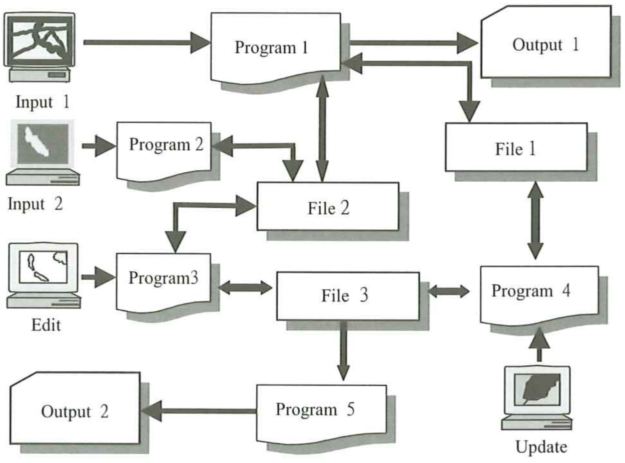

Figure 2. An expanded transaction-based data processing system

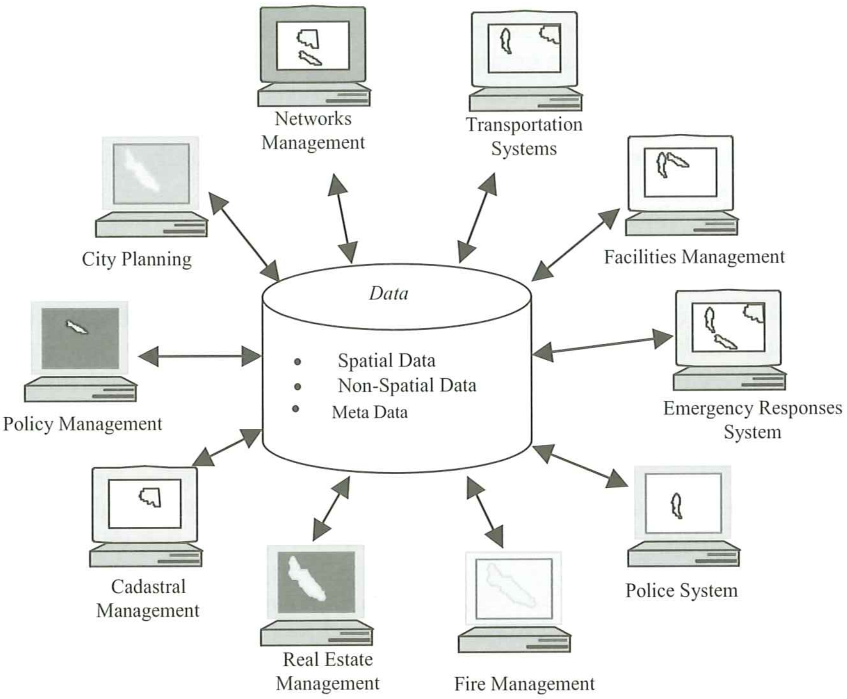

Figure 3. Centralized data repository supporting various data applications

the system allows the users to change the relative importance of factors in analysis, both to evaluate the sensitivity of solutions and to reflect different opinions and objectives of the solutions. Obviously, the operational systems (shown in Figure 4) that are developed for specific purposes are not suitable for providing policy makers or managers with an effective tools on comprehensive data analysis and decision making.

The concept of Decision Support Systems (DSS) or Executive Information Systems (EIS) was proposed to address these issues. In many cases, there is no difference between the DSS and the EIS. However, generally speaking, DSS tend to focus more on details and are targeted towards low to mid-level managers. While, ESS have generally provided a higher level of consolidation and a multi-dimensional view of the data, as high level executives need more the ability to slice and dice the same data than to drill down to review the data detail.

These two similar and overlapping categories are perhaps the closest precursors to the data warehousing systems. Some of the systems have been developed in the past years (Armstrong and Densham, 1990; Bennett, 1983; Fedra ans Reitsma, 1989). A review of the development of spatial decision support system (SDSS) can be found in Densham and Goodchild (1989).

Data warehousing is a repository of large amounts of historical data. It is also a collection of decision support technologies, aimed at enabling the knowledge worker

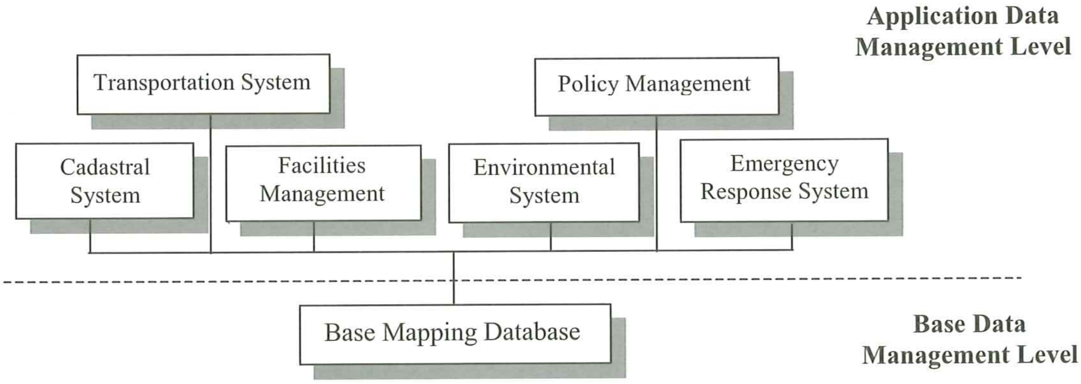

Figure 4. A hierarchy of urban data management systems

(executive, manager, and analyst) to make better and fast decisions. Today's data warehousing systems provide various analytical tools that are not available in their precursors and the design is no longer derived from the specific requirements; and, as we will see later, data warehousing systems are most successful when their design aligns with the overall business structure (Anahory and Murray, 1997; Brackett, 1996; and Lanfond, 1998).

The widely accepted definition of data warehouse is the one provided by Inmon (1992), the father of data warehouse technology. Data warehouse is a subject-oriented, integrated, time-variant, and non-volatile collection of data in support of decision making process. The difference between the data warehouses and the operational databases can be summarized as (Gill and Rao, 1996):

Subject oriented

It organizes and presents data from the perspective of the end user. Most operational systems organize their data form the perspective of the application. The key to the design of the operational systems is performance. Accordingly, data is organized for fast application retrieve. While the data in data warehouses is organized in support of various analysis applications.

Management of large amounts of information including historic data

Most data warehouses contain historic data that is often removed from operational systems because it is no longer needed for operational applications and it may cause the degradation of system performance. An important of features of data warehouses is that it manages an extreme large volume of data with different time, scale and region.

Integration of information from many operational databases

It is in fact that many software applications and databases have been developed for various operations. Data warehouses are needed to collect and organize the data that these applications have gathered over the years. Because of the diversity of storage technologies, database management techniques, and data semantics, integration of various data types, heterogeneous database schemas as well as applications on difference platforms presents a challenging task.

Data warehouses are designed explicitly for decision making not for specific applications. The separation of operational data from the analysis data is the most fundamental data warehousing concept. This separation is supported by several factors such as performance concerns, large volume of data, and difference of data structures and tools used in the data warehouses, etc.

It is worth mentioning that Data Marts are now a popular concept of offering smaller, targeted data warehouses, usually at a lower overhead cost.

III. SPATIAL DATA WAREHOUSES (SDW)

Concept of Spatial Data Warehousing

Geographic or geospatial information systems (GIS) have been the cornerstone of urban information systems. Eighty to ninety percent of all the urban information collected and used is location-related. It would be difficult to think of urban data management systems that do not use geospatial data. With the increasing development of many business data warehouses in organisations, spatially enabling data warehouses have gained of considerable interest recently. It is understandable that spatial data warehousing is a natural extension of data warehousing technology with an emphasis on making use of spatial data. Similar to the business data warehousing, the concept of spatial data warehouses evolved from the needs to store and manage vast amounts of geospatial data and make it readily available for analysis and decision making.

The success of any data warehouse depends on whether it provides the right data to the right user in a timely manner. Spatial data warehouse describes a collection of geospatial data to support spatially related business activities and decision making. It provides a common data model that integrates seamlessly both spatial and non-spatial data. This model enables data filtering, transformation, aggregation, summarization, integration, and deployment. As opposed to operational systems, it is designed to maintain historic data such as vector based geographic features acquired at different time periods. Spatial data warehouse also maintains catalog data, meta-data and administrative data. Data capture information such as method, scale, coverage, accuracy and time period can be documented. Spatial data warehouse not just provides data services, but includes a set of tools that support tabular reporting, spatial query, visualization, on-lined analytic process (OLAP), and data mining.

Spatial data warehouse is evolved in parallel with the data warehousing technology developed for business applications. Its special characteristics can be summarized as,

- being capable of integrating heterogeneous spatial data sources maintained by different GIS software system, supported by different computing platform and stored in different media,

- being capable of handling a variety of data types, vector, raster, matrix, and textural data,
- being capable of inputting or transforming numerous spatial data formats,
- storing historical and time-variant data sets,
- supporting on-line data distribution and access,
- supporting spatial aggregation and generalization,
- supporting sophisticated data display and visualization, and
- supporting spatial data mining and on-line analytic processing.

An Architecture of Spatial Data Warehouse

An architecture of SDW is illustrated in Figure 5. It is a multi-tier environment composed of three main components: data transformation, data management, and data access and usage. The data sources for SDW are either from operational databases or external data sources. They could be spatial vector data, CAD data, imagery, and multimedia data that resides in operational databases, files or networks. These data sources are loaded into the data warehouse through the data transformation component. Data transformation performs data filtering and cleaning, validation, integration and aggregation. The detail data normally found in the operational databases is subject to be aggregated before it enters into the data warehouse.

The data management component is composed of an extended database management system (E-DBMS) to support the capability of storing, managing and access of geodata. One of the core modules in data management is so called spatial data cartridge which is an engine to manage spatial data in relational databases. This design facilitates the operations of spatial data while maintains the interoperability, data security and data integrity. Various spatial data cartridges have been developed in recent years. Some of them are universal, i.e., it can be plugged into other databases and makes the database spatially enabled, for example, ESRI Spatial Data Engine. Others are proprietary. It is designed to expand the capability of the existing databases to handle spatial data, such as Oracle Spatial Cartridge, IBM DB 2 Spatial Extender and Informix Data Blade etc. Meta-data management is also an important module in data warehouses. It provides the complete description of the data stored in the data warehouse. It also describes the pre-defined queries and reports, as well as information regarding the aggregation and summarization of data.

The data access and usage component supports functions such as navigation and browsing, query and analysis, on-line analytic process (OLAP), and data mining. This component is normally built based upon the standard client/server model. This design makes the system more flexible and scalable in terms of

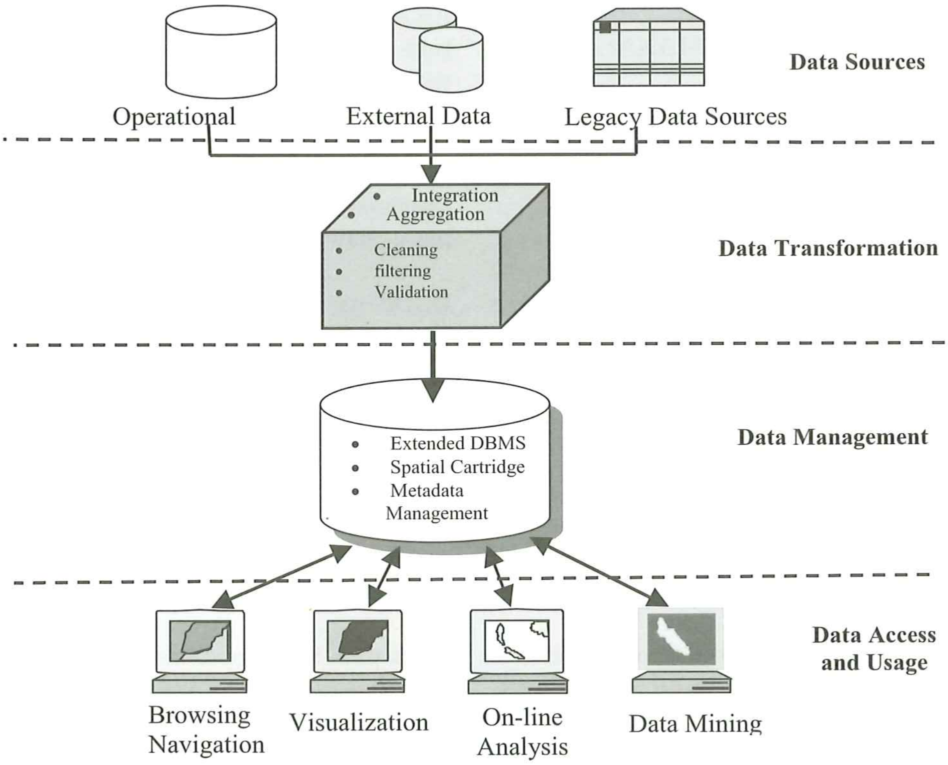

Figure 5. An architecture of a spatial data warehouse

adding more functions and combining more analytic tools. As mentioned, the system that supports decision making should allow users to combine analytical models and data in a flexible manner and enable users to explore the solution space by generating a series of feasible alternatives. It is demanding that the above functions are available through the Internet or Intranet environments.

Among these tools, OLAP and data mining are two primary tools developed for decision making. OLAP is a sophisticated form of query methodology used to aggregate and summarize data in a data warehouse. The basic conceptual model that drives OLAP tools is the multidimensional view of data. OLAP has been researched and used extensively (Kimball, 1996). Excellent surveys of available OLAP tools and vendors can be found in an article by Neil Raden in Barquin and edelstein (1997). Data mining is a more complex query methodology used to discover hidden relationships or trends in the data. There are three fundamental approaches of spatial data mining, namely, classification (supervised), clustering (unsupervised) and visualization. There are various models available for data mining. Besides statistics models, they are decision trees, genetic algorithms, neural networks, agent network, and hybrid models. For more detailed information regarding the data mining methodology, one can refer to books (Groth, 1998 and Cabena et al., 1997).

IV. INTEGRATED URBAN DATA MANAGEMENT FOR DECISION SUPPORT

Needs for An Integrated Urban Data Management Environment

In reviewing the development of data warehousing, one can understand that the fundamental requirements of the operational and analysis systems are different: the operational systems need performance, whereas the analysis systems need flexibility and broad scope. It has rarely been acceptable to have business analysis interfere with and degrade performance of the operational systems. Spatial data warehousing approach provides an optimal solution to the development of a digital environment within which decision makers can explore, structure and solve complex urban problems.

In the next 10 years, a much greater emphasis on informal decision-making will be placed on the development of the urban data management environment. This environment will support the individual and group use of the data resources. It enables the integration of spatial representation, modeling, simulation and planning. It also allows decentralized interaction and decision-making across networks. More importantly, it is accessible by not just technical individuals but also non-technical users. The increasing growth of the Internet is clear evidence of the potential for the kind of environments. It will see that the spatial data warehousing technology will play a key role to this transition.

It has been recognized that the development of large number of diverse operational databases or GISes have contributed largely to the disparate data problem. Managing and archiving the historic data from these systems has been one of challenging problems. With the increasing use of desktop computers and network facilities, it makes the databases and data even more fragmented. SDW promises to integrate the data coming from these operational databases or GISes into a manageable data repository.

SDW can serve not only as an effective platform to merge data from multiple current applications; it can also integrate multiple versions of the same application. For example, a urban transportation system may have migrated to a new standard one that replaces an old mainframe-based, custom-developed legacy application. The data warehouse system can serve as a very powerful and much needed platform to combine the data from the old and the new applications.

An enterprise-wide solution to spatial data integration and sharing is demanded to coordinate management and sharing of disparate data sets between the urban departments, It is this solution that can avoid the duplication of data efforts, making contradictory plans and reduces the inconsistency of data. All of the above requirements lead to the development and building a SDW-based data management system.

Building a Spatially Enabling Data Warehouse for Decision Support

A spatially enabled, integrated, distributed, non-volatile spatial data warehouse should be developed to integrate with the existing two-tiered architecture as shown in Figure 4. This system can be built on top of all existing application systems and forms the third tier in a three-tiered architecture. Figure 6 illustrates this architecture.

In the past, legacy systems archived data onto tapes. As it became inactive, many analysis reports ran from these tapes or mirror data sources to minimize the performance impact on the operational systems. A data warehouse project may start of archiving historic data. The cost of maintaining the data once it is loaded in the data warehouse is minimal. Most of the

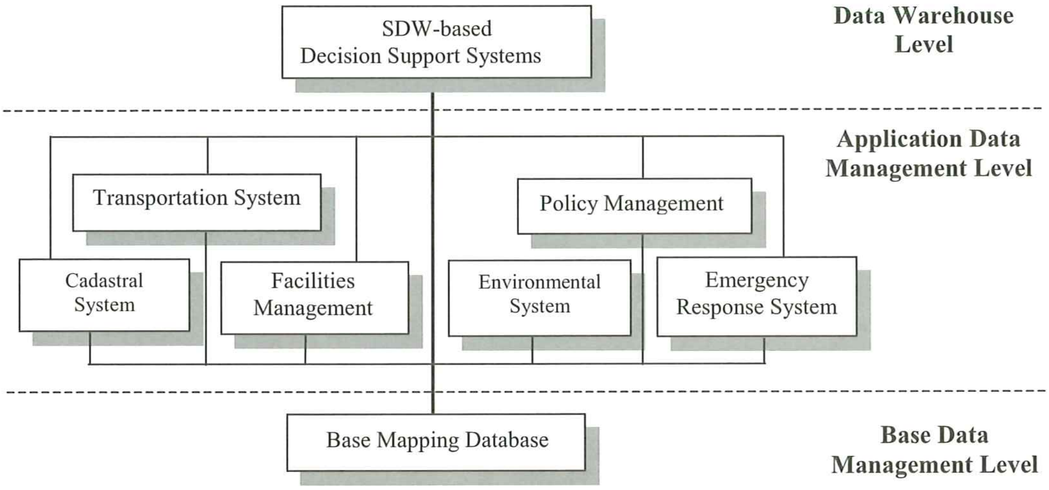

Figure 6. A three-tiered architecture for urban data management

significant costs are incurred in data transfer and data scrubbing. With data warehouses, data can be kept for a very long period. In fact, many data warehouses are built at the same time that the operational applications are deployed.

When building a SDW solution, besides the institutional and management issues, the following technical issues are of particular importance:

Common data model

This model is generic and extensible. It is designed in a way that data from various sources (external sources, operational databases, and various GISes) can be accommodated. Seamless integration of both spatial and non-spatial data must be considered first. The model supports the representation of multiple geometric features such as vector, raster, and matrix. It also allows for different geometric representations to be stored for the same features. Due to the various spatial data structures and spatial operations, the common data model has to be carefully designed to ensure the maximum flexibility and openness. Efficient spatial indexing and partitioning is also required to be accounted in both the logical and physical design of the model.

De-normalization

Normalization is a relational database modeling process where the relations or tables are progressively decomposed into smaller relations to a point where all attributes in a relation are very tightly coupled with the primary key of the relation. Most data modelers try to achieve the "Third Normal Form" with all of the relations before they de-normalize for performance or other reasons. Some of the reasons for de-normalizing the data warehouse model are the same as they would be for an operational system, namely, performance and simplicity. The data normalization in relational databases provides considerable flexibility at the cost of the performance. This performance cost is sharply increased in a spatial data warehousing system because the amount of data involved may be much larger. A three-way join with relatively small tables of an operational system may be acceptable in terms of performance cost, but the join may take unacceptably long time with large tables in the spatial data warehouse.

Summarization, aggregation and generalization

Some non-spatial attributes that are essential to the operational databases are likely to be deemed unnecessary for the spatial data warehouse, and may not be loaded and maintained in the spatial data warehouses. For spatial data, aggregation and generalization is important. The combined use of these techniques will improve the efficiency of data retrieve and enhance on-line data access and spatial visualization. Automated generalization enables the data to be displayed at multiple scales without storing associated multiple scaled data sets. Much more emphasis is being placed on these techniques, as Internet-based applications are becoming more popular.

Meta-data management

Due to the heterogeneous nature of data sources (data, databases and applications), the use of a meta-content repository in SDW has been proposed (Kucera, 1998). The meta-content is comprised of two components: metadata and meta-information. Metadata controls database contents for system analysis while meta-information describes data's fitness for use and database contents for interpretation by end users. The design of meta-content must be compatible with an open approach in order to take advantages of existing and emerging international standards, such as ISO/TC 211.

V. SUMMARY

Inability of GISes and operational databases to support decision making leads to the development of the spatial data warehousing technology. The spatial data warehousing system provides an extensible data environment for the analysis of large amounts of urban data. It is capable of integrating heterogeneous spatial and non-spatial data sets, various applications from different platforms, a variety of databases with different schema into an integrated repository. As discussed, the spatial data warehousing system can be built under the existing two-tiered data environments that are common for most municipalities. Therefore, a three-tiered architecture for integrated urban data management environment can be established. The author accordingly would like to predicate that the spatial data warehousing technology will play a key role in the next 10 years for the development of integrated urban data environments in support of decision making.

ACKNOWLEDGMENTS

The author would like to thank Mr. Chuanyun Fei for his assistance in preparing the figures in this paper.

REFERENCES

[1] Anahory, S., and D. Murray, 1997, *Data Warehousing in the Real World: A practical Guide for Building Decision Support Systems,* Addison-Wesley, ISBN: 0201175193.

[2] Armstrong M. P., and P J Densham, 1990, Database organization alternatives for spatial decision support systems. *International Journal of Geographical Information Systems,* 4:3-20.

[3] Barquin, R., and H. Edelstein, (editors) 1997. *Planning and Designing the Data Warehouse*, Prentics Hall, Upper Saddle River, NJ, Chapter 10 (OLAP).

[4] Bennett, J. L., 1983, *Building Decision Support Systems.* Addison-Wesley.

[5] Brackett, M. H., 1996, *The Data Warehouse Challenge: Time Data Chaos*, Wiley Computer Publishing, ISBN: 0471127442.

[6] Cabena, P., et al., 1997, *Discovering Data Mining: from Concept to Implementation*, Prentice Hall PTR, NJ. ISBN: 0137439806.

[7] Densham, P. J., and M. F. Goodchild, 1989, Spatial decision support systems: a research agenda. *Proceedings of GIS / LIS'89. ACSM,* Bethesda Maryland, pp. 707-16.

[8] Fedra, K., and R. Reitsma, 1989, Decision support and geographical information systems. *Proceedings of the GIS Summer Institute*. Kluwer, Amsterdam.

[9] Gill, H., and P. Rao, 1996, *The Official Client / Server Computing Guide to Data Warehousing*, QUE Corporation, ISBN: 0789707144

[10] Groth, R., 1998, *Data Mining: A Hands-On Approach for Business Professionals*, Prentice Hall PTR, ISBN: 0137564120.

[11] Huxold, W.E. 1991, *An Introduction to urban geographic information systems*. Oxford: Oxford University Press.

[12] Inmon, W. H., 1992. *Building the Data Warehouse*, Wiley-QED Publishing Group, Somerest, NJ.

[13] Kimball, R., 1996. *The Data Warehouse Toolkit*. John Wiley, 1996.

[14] Kucera, H., 1998. Delivering Multidisciplinary Systems, *Proceedings of GIS'98 / RT'98*, Toronto, Canada, April 6-9, 1998, pp.294-298.

[15] Lanfond, P., 1998, Designing and Building the Distributed Geospatial Warehousing Architecture, *Proceedings of GIS'98 / RT'98*, Toronto, Canada, April 6-9, 1998, pp.188-190.