53

ON-LINE ANALYSIS OF UTILITY NETWORKS

Zdeněk Kouba, Kamil Matoušek, Petr Mikšovský
Department of Cybernetics, Czech Technical University in Prague
Technická 2, 166 27 Prague 6, Czech Republic
{kouba, matousek, miksovsp}@labe.felk.cvut.cz

Geographical information systems (GIS) are often used as visualization and analytical means for utility networks applications, because they enable to store information on geographical objects together with their topological and geographical relations. They handle large amounts of geographical data and provide spatial query evaluation on this data. The computationally expensive spatial queries may be improved thanks to the development of on-line analytical processing systems (OLAP) speeding-up the analysis of huge amounts of data stored in large databases. An integration of data warehouse and GIS technologies is the way to enable on-line analysis of geographical information and present its results in geographical contents by native means of the GIS system.

1. INTRODUCTION

Current geographical information systems (GIS) handle large amounts of geographical data that are usually stored in relational databases. From their nature, GIS systems store information on geographical objects together with their topological and geographical relations and they enable spatial query evaluation. As spatial queries are usually very expensive form the computational point of view, major database vendors developed special plug-ins in order to make retrieval of geographical data more efficient. For example, Oracle offers their Spatial Cartridge based on quad-tress, Informix offers the Spatial Data Blade based on R-trees, etc. (Gavrila, 1994, Guttman, 1984). This approach is suitable for those spatial queries, which select objects in certain user-defined area. It does not help so much in the case of analytical queries.

On the other hand we can observe a fast development of so called OLAP systems, i.e. on-line analytical processing systems. The development of such systems has been originally motivated by the need to speed-up the process of analysis of huge amount of data stored in very large databases (Kurz, 1999).

The idea of integration a data warehouse and GIS technologies (Kouba et al., 2000) promises to be the way of enabling on-line analysis of geographical information and present its results in geographical contents by native means of the

V. Mařík et al. (eds.), *Knowledge and Technology Integration in Production and Services*

GIS system. The data warehouse contains materialized views on geographical data. Instead of running a complicated and time consuming spatial query each time when some information is required, the on-line analysis is run on pre-aggregated data. Such integration enables top executives to carry out on-line analytical processing of data originated in a geographical information system. Moreover, it makes possible to present the analysis' results in corresponding geographical context by the native graphical means of the respective geographical information system.

From this perspective the integration is not just a one-way connection. GIS plays a two-fold role in the integrated system. It is not only the data source, from which the data is extracted and pumped into data warehouse. It is also the presentation platform for presenting results of the analyses.

The concept has been tested in a test-bed application oriented on prediction of water consumption in particular nodes of a water supply network.

2. DATA WAREHOUSE

Even a user without special education on data modeling is able to run prepared on-line analysis on a data warehouse. Data warehouse provides him by a transparent concept of a multidimensional abstraction of the data stored in the data warehouse (Kimball, 1996). He understands very well the semantics of data stored in the data cube and the semantics of the corresponding axes (dimensions) of the cube.

He is allowed to carry out very natural analytical operations like slicing the cube, pivoting it, drilling inside the cube etc. These operations are called OLAP (on-line analytical processing) operations.

However, implementing the data cube directly by means of a multidimensional database (MOLAP) is quite rare case. In practice the population of the data cube is rather sparse and therefore the straightforward multidimensional implementation would be inefficient. MOLAP is used for implementation of data marts, i.e. excerpts from the data warehouse with low number of dimensions.

Usually, the data warehouse itself is implemented by means of relational database technology (ROLAP) and is built on the top of a relational database management system. Current state of the art in relational database technology makes possible to process huge amount of data efficiently by means of massive parallel processing.

In relational implementation the OLAP subsystem makes possible to translate the multidimensional queries into SQL.

3. DATA WAREHOUSE MODELING

The data warehouse model is built on several basic concepts. Let us mention shortly and informally some of them.

Any elementary data cell in the above mentioned data cube represent a value of a given fact in context of corresponding positions along the particular axes of the data cube. Each axis represents a dimension of the data warehouse. There is defined a number of aggregation levels for each dimension. It means, each dimension of the data cube can be viewed from different levels of detail. E.g. the fact turnover in a

retail company can be observed from a perspective of weekly turnover in particular district. In this case week is a selected aggregation level of the time dimension, whereas district is the selected aggregation level of the location dimension. By changing the aggregation levels along particular dimensions we change the granularity of the observed data.

Not any two aggregation levels of given dimension may be comparable. For example the aggregation levels month and week of the time dimension are not comparable, as there exists instances of the week aggregation level, which belong to two instances of the month aggregation level. It means that there exists a partial ordering on the set of all aggregation levels of any dimension. This partial ordering may be represented by so-called aggregation graph of given dimension.

The location dimension is the basic concept of GIS – data warehouse integration. It represents the common context of data both in GIS and the data warehouse.

4. INTEGRATION MODULE

The core component of the designed approach is the integration module.

The integration module (IM) has three main functions:

- IM enables the extraction, transformation and load process (ETL) (Kouba et al., 1998) populating the data warehouse by data originating in GIS – see the data stream labeled by (1) in Figure 1.
- IM makes possible to synchronize the status of the GIS with the current status of the data warehouse – see arrow labeled (2) in Figure 1.
- GIS participates in formulation of the multidimensional OLAP query – see arrow labeled (3) in Figure 1.

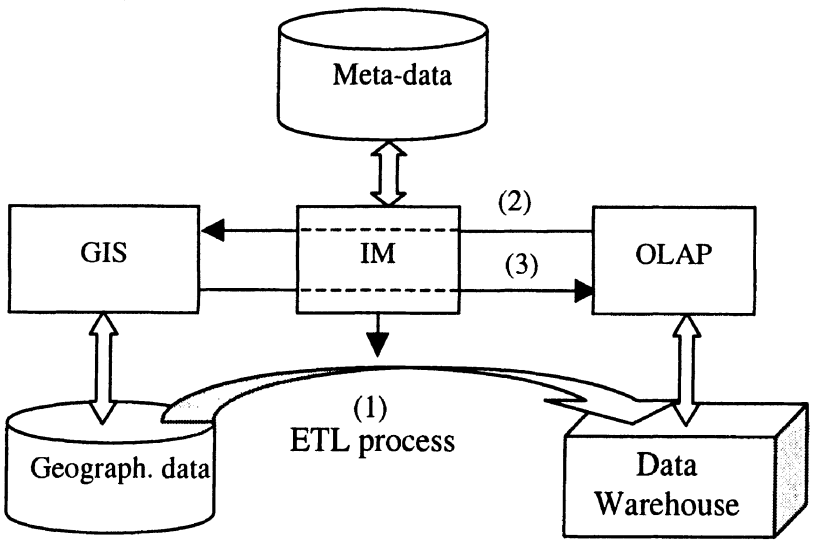

Figure 1 – The Overall Architecture

The integration module is built around the meta-data repository (described bellow), which contains mainly the following types of meta-data:

- Data model of the data warehouse
- Data model of the GIS data source
- Data transformation scripts
- GIS – data warehouse class and instance correspondence
- Definition of data security policies

Purpose of these individual types of information will be described in following paragraphs.

4.1 Meta-Data Repository

Meta-data repository consists of an abstract syntax tree (an internal meta-data representation, which serves as a vehicle for keeping all the meta-data for the system in one place), parser, and methods for meta-data access and maintenance.

ETL operations form an inseparable part of integration module. Thus, the meta-data concept has been enriched by a part related to these ETL operations.

A parser of the meta-data language has been developed. The parser is based on object-oriented version of the PCCTS toolkit by Purdue University. During the parsing process the basic syntax check is performed. The parser builds an internal representation in a form of an abstract syntax tree (AST). All the meta-data is kept in the only tree, which starts from the artificial root. Particular meta-data files are mounted into this tree the size of which is almost unlimited.

Currently, the core of an object-oriented toolkit for the AST manipulation has been finished. It enables the access to the meta-data repository, navigation in the AST, export to text file as well as its parsing into AST. The toolkit is designed as a modular system making possible to add additional modules in future. An interactive graphical editor enabling visualization and maintenance of the AST has been developed as well.

4.2 Extraction, Transformation and Load Process

The extraction, transformation and load process (ETL) is responsible for population the data warehouse from multiple data sources. The designed solution handles unification of data coming from heterogeneous data sources with not fully compatible data models.

Two types of scripts stored in the meta-data repository define the ETL process.

- The transformation scripts extract data from data sources, aggregate them and store the result in the data warehouse.
- The other type of scripts enables evaluation of data validity. Under certain circumstances it is possible to reconstruct wrong or missing data, provided that the data contains some redundancies.

4.3 Class and Instance Correspondence

For purposes of the system integration we consider an object GIS model distinguishing these basic elements:

- *GIS objects* represent individual data items (points, lines, areas etc. with attached sets of data values).
- *GIS classes* are abstract groups of objects of the same level (region, district, building, etc.).

These GIS classes relate to respective aggregation levels of *location* dimension on the data warehouse side. Similarly, the GIS objects correspond to instances of respective aggregation levels on the data warehouse side. Such a correspondence is set up by means of geographical meta-data class defined in the meta-data repository.

It is not necessary to map each internal GIS class to geographical meta-data class and vice-versa. In general, the GIS system should be able to manipulate with the whole geographical data under examination. For example, the GIS system need not be aware of the details about the aggregation graph of the *location* dimension of the data warehouse.

In most cases we expect the GIS system to be a geographical data source providing the rest of the system with geographical data. A subset of this data is then retrieved and offered to the data warehouse.

The binding elements between GIS and data warehouse are the elements of the data warehouse *location* dimension and the GIS taxonomy objects. The task of the integration is to provide three kinds of necessary dynamic correspondences:

1. *Class correspondence* maps particular aggregation levels of the location dimension to the corresponding GIS taxonomy levels and vice versa. This is relatively long-time static pre-defined information stored in integration meta-data.
2. *Instance correspondence* maps particular instances of aggregation levels to the instances of the "classes" in sense of the previous item and vice versa. This is a more dynamic part of the integration information and it guaranties the run-time data integrity. Integration module should keep track of the instance changes in both GIS and DWH and propagate them to the second sub-system.
3. *Action correspondence* is the most dynamic correspondence, which ensures navigation consistency. E.g. after changing aggregation level using particular front-end tool, level information has to be changed in meta-data, data warehouse and GIS. Another example: When performing a GIS selection, just particular sub-cube should be considered for calculation, i.e. corresponding filter should be applied in the data warehouse and the front end tool should show it.

Meta-data objects support all of the three correspondences.

Taxonomy structure of GIS objects is coupled with aggregation levels of a data warehouse via class correspondence. Correct mapping of particular objects is guarded by class correspondence (GIS object "Czech Republic" is related to the instance "Czech Republic" of an aggregation level *country* of the *location* dimension of the data warehouse).

4.4 Data Security Aspects

The issues of authorization and access control concepts relating security concepts to data warehouse technology have been studied. In contrast to record-oriented access policy in typical database systems, in case of coupling both data warehouses and GIS the access rights should be related to the data granularity.

A user may exist, for which the data on very detailed level of granularity are not accessible. Such data may have private nature and therefore it has a restricted access. On the other hand, some data on very general granularity level may not be available to the user, as the data is of strategic nature.

The new cell-level security developments are useful for the needs of GIS-DWH interoperability.

5. APPLICATIONS

A pilot project: "Drinking water distribution system in western Bohemia" use the designed approach. It runs on commercial GIS and Data Warehousing systems to assure the concept, its usefulness and its extensibility by other commercial or non-commercial information systems. Examples of these applications are presented here.

Extensible Markup Language (XML) was determined as a meta-language for data transfer. Selected XML documents contain the particular information that is necessary for inter-system communication and synchronization. Various data models have to be mapped to each other using both static and dynamic portions of such meta-data. Document Object Model (XML-DOM) enables easy document and data transformation among different software platforms. The implementation of the Integration Module (Matoušek et al., 2001) was designed using modern technologies like Component Object Model (COM), Office Web Components (PivotTable Service) and ActiveX Data Objects (ADO). End users can use a broad variety of front-end tools to manipulate and analyze the geographical and dimensional data of their interest. The applicable user interfaces include Dimensional Navigator Snap-In for GIS, Microsoft Excel Pivot Table and others. The tested GIS system is ArcView 3.2.

5.1 Drinking Water Distribution System in Western Bohemia

This pilot application concerns drinking water distribution and consumption. The distribution network begins in the manufacturing part, where the natural water quality is improved for the water to become drinkable and then water is pumped to the primary water supply. Water is distributed to customers indirectly via storage reservoirs. The best quality of water and most efficient production is achieved by constant amount of water produced per time unit. Therefore the consumption peaks must be foreseen and the reservoirs must be pre-filled to contain enough water for the following peak. However, keeping non-necessarily high level of water in water reservoirs is bad, as the water quality is affected by slow exchange of reservoir contents.

The GIS and data warehouse models correspond to each other in following "classes" (levels) of geographical dimension:

- *Reservoir* - represents a water reservoir containing one or more tanks
- *Pipe line* - corresponds to a pipe line connecting several reservoirs
- *Region* - defines an administrative region.

A dimensional data model of the drinking water distribution application, which covers all user requirements, is shown in Figure 2. The model consists of one fact table, two dimensions (i.e. time and geography), and three look-up tables (weather, area, and pipe line).

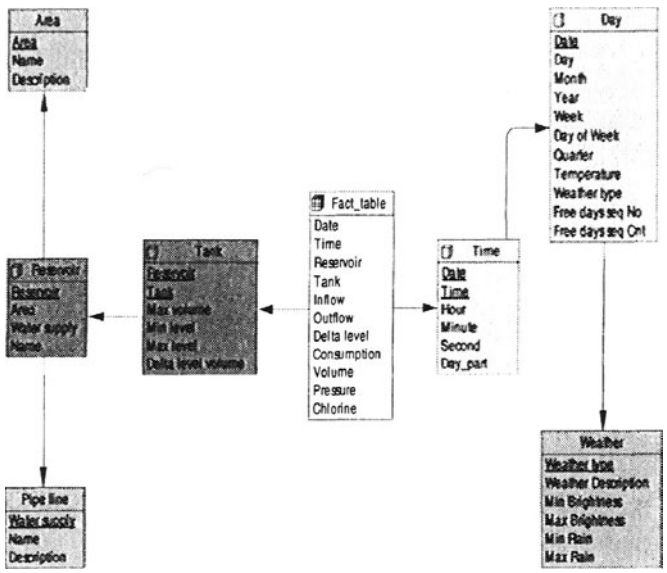

Figure 2 – Water Distribution Data Model

7. CONCLUSIONS

The basic principles for integration of geographical information systems with data warehouses using integration module were formulated. The two integration prototypes concerning two pilot applications were described.

Data warehouses represent a technology, which is able to analyze huge amounts of data in relatively short time. The OLAP engine is optimized for fast query evaluation. Data warehouse principle is based on a specific multidimensional data model, which is powerful in performance and easily understandable. This power is augmented by pre-calculated aggregations of some subsets of the data on various dimensional levels that radically increase the performance.

On the other hand, geographical information systems are designed for storage of structured and spatial information with all the spatial data specific features and functionality. Its drawbacks are the spatial queries that are computationally very demanding.

Therefore GIS integrated together with a data warehouse technologies profit from each other. For example data warehouses can store the pre-calculated results of spatial queries posed within GIS. The conventional OLAP and GIS coupling is capable to analyze on-line pre-defined geographical areas.

The concepts can be used on different system platforms, are extensible and open for further research activities.

8. ACKNOWLEDGMENTS

The work related to this paper has been carried out with support of the INCO-COPERNICUS No. 977091 research project *GOAL – Geographical Information On-line Analysis* and the research grant of the Czech Ministry of Education, Youth and Sport: "Decision Making and Control for Industrial Practice", No. 212300013.

The authors want to express thanks to their colleagues from the *Gerstner Laboratory for Intelligent Decision Making and Control* for creating a friendly environment.

9. REFERENCES

1. Gavrila, DM. R-tree Index Optimization, CAR-TR-718, Comp. Vision Laboratory Center for Automation Research, University of Maryland, 1994
2. Guttman, A. "R-trees: a dynamic index structure for spatial indexing", In Proc. of SIGMOD Int. Conf. on Management of Data, 1984
3. Kimball R. The Data Warehouse Toolkit. Practical Techniques for Building Dimensional Data Warehouses, John Willey & Sons Inc., 1996
4. Kouba Z, Mařík V, Mikšovský P, Tjoa AM. "Data Warehousing and Geographical Information", In Proc. of 4th World Multiconference on Systemics, Cybernetics and Informatics - SCI 2000, 2000
5. Kouba Z, Matoušek K, Mikšovský P, Štěpánková O. "On Updating the Data Warehouse from Multiple Data Sources". In DEXA '98, Vienna, Springer-Verlag, 1998
6. Kurz A. Data Warehousing – Enabling Technology, (in German), MITP-Verlag GmbH, 1999
7. Matoušek K, Mordačík J, Janků L. "On Implementing the Data Warehouse – GIS Integration". In Proc. of World Multiconference on Systemics, Cybernetics and Informatics - SCI 2001