# A DATA WAREHOUSE ARCHITECTURE FOR INTEGRATING FIELD-BASED DATA

Alberto Salguero, Francisco Araque and Ramón Carrasco
*Dpt. LSI – ETSII, University of Granada (Andalucía), España*

Keywords: Spatial Data Warehouse, extraction, field-based data.

Abstract: Spatial DataWarehouses (SDWs) combine DWs and Spatial Data Bases (SDBs) for managing significant amounts of historical data that include spatial location. Some spatial information can be seen as a continuous field, and the information of interest is obtained at each point of a space. The previously proposed extensions of the multidimensional data model, used in Data Warehousing, only deal with spatial objects. None of them consider field-based information. This paper presents a Data Warehouse architecture that automatically determines the best parameters for refreshing and integrating field-based data from different data sources.

## 1 INTRODUCTION

Spatial Data Bases (SDB) have a long experience in managing spatial data, and there is extensive research referring to spatial index structures, storage management, and dynamic query formulation. In order to manipulate spatial objects, a SDB must include special data types to represent geometric characteristics of objects. On the other hand, space also can be seen as a continuous field, and the information of interest, temperature, pressure, elevation…, is obtained at each point of the space. Geographic Information Systems (GIS) are the obvious potential candidates for such tasks. While having some spatio-temporal analytical capabilities, it is recognized that existing GISs per se are not adequate for decision-support applications when used alone (Bédard et al., 2001). Among the possible solutions, the coupling of spatial and non-spatial technologies, GIS and OnLine Analytical Processing (OLAP) for instance, may be an interesting option.

OLAP systems are usually implemented through Data Warehouses (DWs). A Spatial Data Warehouse (SDW) combines DWs and Spatial Data Bases for managing significant amounts of historical data that include spatial location.

Inmon (Inmon, 2002) defined a DW as "a subject-oriented, integrated, time-variant, non-volatile collection of data in support of management's decision-making process." A DW is a database that stores a copy of operational data with an optimized structure for query and analysis. In terms of a more limited scope, a new concept is defined: a Data Mart (DM) is a highly focused DW covering a single department or subject area.

The generic architecture of a DW is illustrated in Figure 1. It can be seen that data sources include existing operational databases and flat files (i.e., spreadsheets or text files) in combination with external databases. The data are extracted from the sources and then loaded into the DW using various data loaders and other tools (Araque and Samos, 2003). The warehouse is then used to populate the various subject (or process) oriented data marts and OLAP servers. Data marts are subsets of a DW categorized according to functional areas depending on the domain (problem area being addressed) and OLAP servers are software tools that help a user to prepare data for analysis, query processing, reporting and data mining.

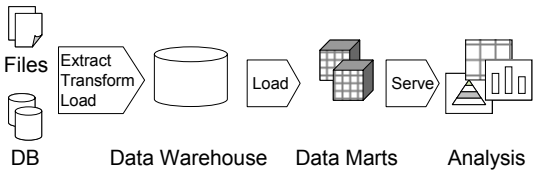


Figure 1: A generic DW architecture.

DWs are usually based on a multidimensional data model to make easy the querying and the analysis of the data. Several extensions of the



Salguero A., Araque F. and Carrasco R. (2007).
A DATA WAREHOUSE ARCHITECTURE FOR INTEGRATING FIELD-BASED DATA.
In *Proceedings of the Ninth International Conference on Enterprise Information Systems - DISI*, pages 577-580
DOI: 10.5220/0002391105770580

multidimensional data model that consider the spatial aspects of data have been proposed last years. These works usually incorporate the concepts of SDBs to the multidimensional data model (Bimonte et al., 2005), (Malinowski and Zimányi, 2004), (Gascueña et al., 2006). These models focus on the representation of complex spatial objects (SO). None of the reviewed models deal with field-based information. The addition of this kind of information to the conceptual data model is not a difficult task. The main problem is that the data should be obtained from autonomous data sources. Usually, the field-based information is created by interpolating the values of several sensors distributed along a surface. In Data Warehousing it is necessary to integrate this semantically-related data from several data sources. Furthermore, when the information refers to the same spatial location it can be consider information semantically equivalent. It is possible that two data sources refer to the same real-world entity at the same time with different spatial or temporal detail.

This paper presents an architecture that automatically determines the best parameters for refreshing and integrating field-based data from different data sources.

The remainder of this paper is organized as follows. In section 2 our architecture for integrating spatial data is presented; in section 3 an illustrative example is explained; finally, section 4 summarizes the conclusions of this paper and discusses future work.

## 2 ARCHITECTURE

The architecture described here can determine the best parameters for refreshing data sources which contain spatial data semantically related. If we use a set of sensors distributed along a surface, for instance, to know the temperature at each point of that surface (interpolating), we need to know the values of all sensors at a given instant in order to make comparable captures of field-based data.

Taking paper (Sheth and Larson, 1990) as point of departure, we propose the following reference architecture. Three more different levels should be considered:

- Component Scheme ST: the conversion of a Native Scheme to our Canonical Data Model (CDM), enriched so that temporal and spatial concepts could be expressed.
- Exportation Scheme ST: it represents the part of a component scheme which is available for the DW designer. It is expressed in the same CDM as the Component Scheme.
- Data Warehouse Scheme: it corresponds to the integration of multiple Exportation Schemes ST according to the design needs expressed in an enriched CDM so that temporal concepts could be expressed

We will describe now the functional architecture in a detailed way.

**Native Schema**. Initially we have the different data source schemes expressed in its native schemes. Each data source will have, a scheme, the data inherent to the source and the metadata of its scheme. In the metadata we will have huge temporal and spatial information about the source: temporal and spatial data on the scheme, metadata on availability of the source…

The temporal parameters we consider of interest for the integration process are (Araque et al., 2006a):

- Availability Window (AW). Period of time in which the data source can be accessed.
- Extraction Time (ET). Period of time taken by the monitoring program to extract significant data from the source.
- Granularity (Gr). It is the extent to which a system contains discrete components of ever-smaller size. In our case it is common to deal with granules like meter, kilometre…

To solve the problem of getting this information from Web accessible sources, we present tools to define and generate wrappers. We define a wrapper interface to specify the capability of Web Sources and extend a wrapper generation toolkit. We use DETC (Data Extraction with Temporal Constraints) to refer to the software tool.

**Preintegration**. In the Preintegration phase, the semantic enrichment of the data source native schemes is made by the conversion processor. In addition, the data source spatial and temporal metadata are used to enrich the data source scheme with temporal properties. We obtain the component scheme (CS) expressed in the CDM, in our case, ODMG-ST (ODMG enriched with spatial and temporal elements).

**Componet and Export Schemas**. From the CS expressed in ODMG-ST, the negotiation processor generates the export schemes (ES) expressed in ODMG-ST. These ES are the part of the CS that is considered necessary for its integration in the DW. For privacy reasons part of the CS can be hidden.

**Integration**. From many data sources ES, the DW scheme is constructed (expressed in ODMG-ST). This process is made by the Integration Processor that suggests how to integrate the Export

Schemes helping to solve semantic heterogeneities (out of the scope of this paper). In the definition of the DW scheme, the DW Administrator participates in order to contemplate the characteristics of structuring and storage of the data in the DW.

Three modules have been added to the reference architecture in order to carry out the integration of the data, considering the data extraction method used:

- *The Temporal Integration Processor* uses the set of semantic relations and the conformed schemes obtained during the detection phase of similarities. As a result, we obtain data in form of rules about the integration possibilities existing between the originating data from the data sources (minimum granularity...). This information is kept in the Temporal Metadata Warehouse. In addition, as a result of the Temporal Integration process, a set of mapping functions is obtained.

- *The Spatial Integration Processor* does the necessary transformations of the spatial data in data sources in order to integrate them. It is necessary to convert all of data to the same format and unit of measurement. It is also responsible of dealing with the different spatial granularity of data we can find in different data sources.

- *The Metadata Refreshment Generator* determines the most suitable parameters to carry out the refreshment of data in the DW scheme. The DW scheme is generated in the resolution phase of the methodology of integration of schemes of data. It is in this second phase where, from the minimum requirements generated by the temporal integration and stored in the Temporal Metadata warehouse, the DW designer fixes the refreshment parameters. As result, the DW scheme is obtained along with the Refreshment Metadata necessary to update the former according to the data extraction method and other temporal properties of a concrete data source.

**Data Warehouse Refreshment**. After temporal integration and once the DW scheme is obtained, its maintenance and update will be necessary. This function is carried out by the DW Refreshment Processor. Taking both the minimum requirements that are due to fulfill the requirements to carry out integration between two data of different data sources (obtained by means of the Temporal Integration module) and the integrated scheme (obtained by the resolution module) the refreshment parameters of the data stored in the DW will be adjusted.

# 3 EXAMPLE

A Decision Support System (DSS) being based on a DW is presented as an example (fig. 2). This can be offered by Small and Medium-Sized Enterprises (SMEs) as a plus for adventure tourism. Here, a DSS is used to assist novel and expert pilots in the decision-making process for a soaring trip (Araque et al., 2006b). These pilots depend to a large extent on meteorological conditions to carry out their activity and an important part of the system is responsible for handling this information. Two web data sources are mainly used to obtain this kind of information:

- The US National Weather Service Website. We can access weather measurements (temperature, pressure, humidity, etc) in every airport in the world. In Spain we can find 48 airports where we can extract this information.
- In order to obtain a more detailed analysis and to select the best zone to fly, pilots can access to the Spanish National Weather Institute (INM) website. There are 205 meteorological stations distributed along the Spanish surface. They are usually refreshed every thirty minutes.

The continuous integration of Web data sources may result in a collapse of the resources of the SMEs, which are not designed to support the laborious task of maintaining a DW up to date.

In our approach, the DW administrator introduces the data sources temporal properties in *DETC* tool and selects the parameters to integrate, for example the temperature. This tool is able to determine the maximum level of detail (granularity).

We find out that in the second source, the information about the temperature can be precise with a detail of "minute" (for example, that at 14 hours and 30 minutes there were a temperature of 15ºC), whereas in the first case it talks about the temperature with a detail of "hour" (for example, that at 14 hours there were 15ºC).

It can also determine the time intervals in which this information is available to be queried (useful when dealing with other kind of data sources).

Applying the temporal algorithms, out of the scope of this paper, we would obtain all possible instants of querying which both sources are accessible at, so the extraction and integration process can be performed (Araque et al., 2006a).

Let us suppose that both data sources in this example are always available for querying. The DWA, who usually wants to get the most detailed information, would select to extract the changes from the first data source every hour and every half an hour in the case of the second one. There is a waste of resources in this approach.

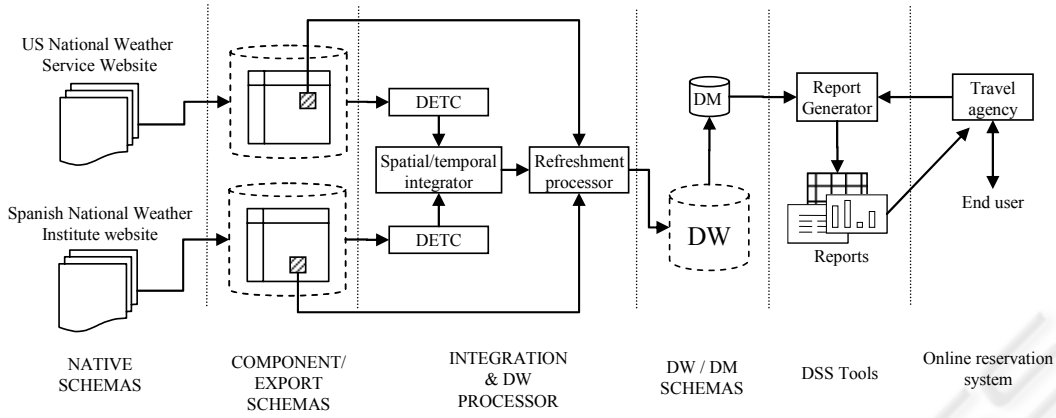


Figure 2: Illustrative example.

The field-based data distribution obtained using only the second data source (every thirty minutes) is not always comparable with the data distribution obtained combining both data sources because almost 20% of the sensors are unavailable (airports sensors are refreshed every hour). Therefore, if we cannot make use of the information obtained from the second data source for analysis when there are not values to get from the first one, we are actually discarding about 205/(205+205+48) = 45% of extracting, transforming and loading processes.

The architecture presented in this work is able to detect this kind of issues and determine that it has no sense to query the second data source every thirty minutes.

## 4 CONCLUSIONS

Several extensions of the multidimensional data model, which is mainly used in Data Warehousing, that consider the spatial aspects of data have been proposed. The problem is that in Data Warehousing it is necessary to integrate data from several data sources. This issue has not been widely studied, especially in the case of field-based data. In Data Warehousing, the integration of this kind of data requires all values of sensors to be accessible in order to compare the results.

This paper presents a Data Warehouse architecture that automatically determines the best parameters for refreshing and integrating field-based data from different data sources.

This work has been supported by the Spanish Research Program under project TIN2005-09098-C05-03 and the by Andalucía Research Program under project 2006/282916.

## REFERENCES

Araque, F., Salguero, A., Abad, M.M., 2006b. Application of data warehouse and Decision Support System in Soaring site recommendation. Proc. Information and Communication Technologies in Tourism, ENTER. Springer Verlag, Lausanne, Switzerland.

Araque, F., Salguero, A.G., Delgado, C., Garvi, E., Samos, J., 2006a. Algorithms for integrating temporal properties of data in Data Warehouse. 8th International Conference on Enterprise Information Systems (ICEIS). Paphos, Cyprus.

Araque, F., Samos, J., 2003. Data warehouse refreshment maintaining temporal consistency. 5th Intern. Conference on Enterprise Information Systems, ICEIS. Angers. France.

Bédard, Y., T. Merrett and J. Han. 2001. Fundamentals of spatial data warehousing for geographic knowledge discovery. Geographic Data Mining and Knowledge Discovery. Ed. H. Miller & J. Han, Taylor & Francis.

Bimonte, S., Tchounikine, A., Miquel, M., 2005. Towards a Spatial Multidimensional Model , DOLAP05, ACM Eighth International Workshop on Data Warehousing and OLAP , Bremen, Germany.

Gascueña, C. M., Cuadra, C., Martínez, P. A., 2006. Multidimensional approach to the representation of the spatio-temporal multi-granularity. In proceedings of the Intern. Conference on Enterprise Information Systems (ICEIS). Paphos, Cyprus.

Inmon, W.H., 2002. Building the Data Warehouse. John Wiley.

Malinowski, E., Zimányi, E., 2004. Representing spatiality in a conceptual multidimensional model. In proceedings of the 12th annual ACM International workshop on Geographic information systems. Washington DC, New York, USA.

Sheth, A., Larson, J., 1990. Federated Database Systems for Managing Distributed, Heterogeneous and Autonomous Databases. ACM Computing Surveys, Vol. 22, No. 3