Available online at www.sciencedirect.com
SCIENCE DIRECT®

Environmental Modelling & Software 21 (2006) 1579–1586

Environmental
Modelling & Software

www.elsevier.com/locate/envsoft

# Database architectures: Current trends and their relationships to environmental data management

Jaroslav Pokorný*

*Charles University, Faculty of Mathematics and Physics, Department of Software Engineering, Malostranské nám. 25, 118 00 Praha, Czech Republic*

Received 11 November 2005
Available online 27 June 2006

**Abstract**

Ever increasing environmental information demands from customers, authorities, and governmental organizations as well as new business control functions are implemented and integrated to environmental information management systems (EIMSs). These systems are often based on traditional file techniques or, more recently, on commercial database management systems (DBMSs). With a production of huge data sets and their processing in real-time applications, the needs for environmental data management have grown significantly. Numerous examples from practice of EIMSs prove that the architecture of DBMS should be open for a permanent evolution. Current trends in database development and an associated research meet these challenges. New information and communication technologies and techniques influence today's DBMSs. They include, among other things, sensor networks, stream processing, processing uncertain and imprecise data, knowledge discovery and intelligent data analysis, as well as wireless broadcast and mobile computing. Both research and practice indicate that the traditional universal DBMS architecture hardly satisfies these trends and new solutions are needed. Rather separate specialized engines connected into networks are beneficial. The paper discusses recent advances in database technologies and attempts to highlight them with respect to requirements of EIMSs.


*Keywords:* Environmental management system; Database management system; Sensor; Sensor network; Stream processing; Uncertain and imprecise data; Knowledge discovery and intelligent data analysis; Wireless broadcast; Mobile computing

## 1. Introduction

Without doubt the world of data is changing, particularly, the nature and sources of information. All these changes have a significant influence on database needs, and consequently, on questions where the database field is and where it should be going. Abiteboul et al. (2005) in their report emphasize two main driving forces today: Internet and particular sciences, like the physical sciences, biological sciences, medicine, and engineering. These sciences produce large and complex data sets that require more advanced database support than current products provide.

Another trend existing since the 1960s concerns the industries having faced ever increasing environmental demands from customers, authorities and governmental organizations. Recently, reflecting these demands, new business control functions are integrated to *environmental management systems*[1] (EMS). For their computerized part we can use the term *environmental information system* (EIS), if we address public environmental information systems, or *environmental management information system* (EMIS), if we deal with industrial environmental information systems. As data or information processing is primarily what we focus on, we will use the term EIS through the paper.

An important observation is that, similarly to the sciences mentioned, EISs process also huge data sets, often continually

* Fax: +420 221914323.
*E-mail address:* pokorny@ksi.ms.mff.cuni.cz

1364-8152/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.envsoft.2006.05.004

[1] By LCA (2005) an EMS is a part of the overall management system that includes organisational structure, planning activities, responsibilities, practices, procedures, processes and resources for developing, implementing, achieving, reviewing, and maintaining the environmental policy.

and with triggering various control actions. Consequently, the needs for environmental data management have grown significantly.

Considering environmental data sets combined with, e.g., business data, emails, documentations etc., adequate information integration mechanisms are needed. Since their beginning the databases have had an integrative role in the world of data. Reuter (2005) argues that the technological evolution of database technology makes database systems even the ideal candidate for integrating all types of objects that need persistence as well as for supporting all the different types of execution that are characteristic of the various application classes.

The most important part of each management system deals with data through querying. When users want to search and use environmental information, the following problems occur (Tomasic and Simon, 1997):

(1) Data do not exist or are insufficient; sometimes this may require synthesis or reproduction of data.
(2) Data are not referenced by data suppliers and therefore hard to locate, or data are referenced under specific classification criteria that are domain-specific.
(3) Data are hard to access; they are either private or of a too high cost, or requiring costly pre-processing (e.g., data must be re-entered manually from paper documentation) or format translation.
(4) Accessed data sets are hard to use because they are inconsistent or non-compatible; for example, access to long time series but standard data collection techniques have not been applied, thereby making adjacent time series not compatible.
(5) The quality of retrieved data is hard to assess; it is often hard to compare data produced using different scientific models because of a lack of documentation about the underlying computational processes.

The database community focuses on information storage, organization, management, and access in software architectures called *database management systems* (DBMSs). Always it is driven by new applications, technology trends, new synergies with related fields, and innovation within the field itself. The problems (1)–(5) are a natural part of today's database research and development. A natural idea is that EISs based on advanced database technologies could help to deal with these issues.

Several technological aspects influence DBMS development. Focusing on the scientific data, it is often coming in streams. The sensor networks producing the data consist of very large numbers of low-cost devices, each of which is a data source, measuring some quantity, e.g. the object's location, or the ambient temperature. Processing such data is usually completely different from the data stored in enterprise databases. Data arrive in high-speed streams, and queries over those streams need to be processed in an online fashion to enable real-time responses. Moreover, in comparison to enterprise data processing, these data are uncertain or imprecise. Other aspects of such data processing include unclear formulation of queries based on common techniques as they are used for example in classical databases. Often we are not able to formulate a query, e.g. in SQL, and despite of the fact we believe on the other hand that something interesting is hidden in our data. In such situations a lack of semantics is apparent. To describe data semantics, metadata and its formal description are necessary.

Data in collections considered creates an ideal platform for using knowledge discovery methods and/or intelligent data analysis. Also online analytical processing (OLAP), data warehouses (DW), and data mining (DM) techniques can help in this context.

The purpose of the paper is to present the main challenges influencing today's database development with respect to the processing environmental data. First, in Section 2 we discuss properties of new data sources. In Section 3 we repeat the concepts of the classical centralized DBMS architecture, as it exists from early 1980s. According to Härder and Reuter (1983), the architecture models five-level abstraction hierarchy. Its implementation has five technological layers that allow to separate some problems and their solutions in relatively independent way. The main part of the paper presents five new technologies influencing database architectures in Section 4. They include sensor data and sensor networks, stream processing, approaching uncertain and imprecise data, knowledge discovery methods and intelligent data analysis, and wireless broadcast and mobile computing. In Section 5, we argue that new DBMS architectures are needed, describing briefly some of their proposals, and give several examples of their occurrences in practice. In conclusions we summarize the basic ideas given in the paper and add a number of other issues that can influence processing environmental data.

## 2. New data sources

Usual enterprise data stored in databases are structured and can be described by a so-called (*database*) *schema*. Such a schema is almost fixed or it is changed only rarely. It is not the case of collections of scientific as well as environmental data. By Reuter (2005), the key properties of these data collections (irrespective of the many differences) are the following:

- The raw data is written once and never changes again. As a matter of fact, some scientific organizations require for all projects they support that any data that influences the published results of the project be kept available for an extended period of time, typically around 15 years.
- Raw data comes in as streams with high throughput (hundreds of MB/s), depending on the sensor devices. The streams have to be recorded as they come in, because in most cases there is no way of repeating the measurement.
- For the majority of applications, the raw data is not interesting. What the users need are aggregates, derived values, or – in case of text fields – some kind of abstract of "what the text says".