Database architectures: current trends and their relationships to environmental data management

Jaroslav Pokorný[1],

Abstract

Ever increasing environmental demands from customers, authorities and governmental organizations as well as new business control functions are integrated to environmental management systems (EMSs). With a production of huge data sets and their processing in real-time applications, the needs for environmental data management have grown significantly. Current trends in database development and an associated research meet these challenges. The paper discusses recent advances in database technologies and attempts to highlight them with respect to requirements of EMSs.

1. Introduction

Without doubts 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, S. 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 datasets that require more advanced database support than current products provide. They also need information integration mechanisms.

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*[2] (*EMS)*. Consequently, the needs for environmental data management have grown significantly.

When users want to search and use environmental information, the following problems occur (Tomasic/Simon 1997): (1) Data do not exist or are insufficient; sometimes this may require synthesis or reproduction of data. (2) Data is not referenced by data suppliers and therefore hard to locate, or data is referenced under specific classification criteria that are domain-specific. (3) Data is hard to access: 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 (e.g. 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 process.

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

[1] Charles University, Faculty of Mathematics and Physics, Malostranské nám. 25, 118 00 Praha, Czech Republic

[2] 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.

problems (1)-(5) are a natural part of today's database research and development. EMSs based on advanced database technologies could help to deal with these issues. Some attempts to influence a development of EMSs by database specialists exist even from the past. For example, the Sequoia 2000 (Stonebraker 1994) project speaks about collaboration between computer scientists and environmental researchers to design a next-generation information system for managing data for global change research.

Obviously, it is not surprising that in many cases only traditional file-oriented solutions are at disposal. For example, the CORIE (Columbia River Estuary) system based on three forms of data (science data, catalogue data, and task data), produces in its simulations 5GB of forecasted data each day (Bright/Maier 2005). Nevertheless, its Metadata Repository is schema-less, no file formats, database access libraries, or XML schemas need be agreed upon. In connection with Internet, web services, and EMS, such solutions seem to be unsustainable.

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 of such data is usually completely different from the data stored in enterprise databases. Data arrives 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 business data processing, this data is uncertain or imprecise. Other aspects of such data include unclear formulation of queries based on common techniques as 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. Also online analytical processing (OLAP) and data mining 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. The rest of paper briefly discusses in sections 2.-5. sensor data and sensor networks, stream processing, approaching uncertain and imprecise data, data mining, and wireless broadcast and mobile computing. In conclusions, we argue that new DBMS architectures are needed, and describe briefly their characteristics.

2. Sensor data and sensor networks

For environmental data a number of new technologies are relevant: inexpensive micro sensor technology that will enable most objects to report their temperature, pressure, state or location, e.g. via a global positioning system, in real time. This information will support applications whose main purpose is to monitor the objects attributes.

Sensor networks provide important data sources and create new data management requirements. For instance, these sensors are generally self powered, wireless devices. Such a device draws far more power when communicating than when computing. Thus, when querying the information in the network as a whole, it is often preferable to distribute as much of the computation as possible to the individual nodes. In fact, the network becomes a new kind of database machine, whose optimal use requires operations to be pushed as close to the data as possible. In a more complicated case, sensors and/or users can be mobile.

Sensor information processing raises many of the most interesting database issues in a new environment, with a new set of constraints and opportunities. Huge datasets, e.g. of environmental data, generated by sensors will be distributed throughout the world, and can come and go dynamically. In other words, sensors can produce continuous, possibly infinite, streams of data.

From one perspective sensor networks are similar to distributed databases, but with inherent real-time properties. One important difference is that the evaluation rate of data produced in a sensor network is much higher that typically considered in distributed DBMSs. This breaks the traditional information inte-

gration paradigm, since there is no practical way to extract and load data into a common database to each such occurrence. Also strategies of query optimization and query processing must be redefined.

3. Stream processing

Environmental data management based exclusively only on the traditional store-and-query model can not handle the volume and velocity of streaming data, whose values might exist a moment. Traditional DBMSs are unsuited to deal with such streams for various reasons (Amato et al. 2004):

- sensor nodes produce and deliver data continuously without receiving requests for that data,
- queries over collected data can be less frequent than data insertions,
- produced data has often to be processed in real time because it can represent events, that need a rapid answer,
- queries run continuously because data streams never terminate, so, they can see system conditions change during their execution,
- because of storage constraints, an entire stream can not be stored in the disk,
- because data streams are possibly infinite, only non-blocking operators can be used,
- if the tuple is not avaiable, so, operators must process data only when nodes make it available.

In consequence, so called *data stream processing systems* have emerged; see e.g. (Carney et al. 2002). A *stream-processing engine* (SPE) is an example of a new data base architecture that enables the execution of queries, computations, and actions on streaming data in real time. Such SPE should accept SQL-like queries, stream oriented, continuous queries and execute them over live event streams with outputting results in real time. In SPEs most of processing is produced in main memory, read or write operations to storage are optional and can be handled asynchronously in many cases.

For example, in a recent pilot program, Streambase developed by Stonebraker (StreamBase Systems, Inc., 2005) should be able to analyze 140,000 messages per second, while a leading relational database could handle only 900 messages per second.

4. Approaching uncertain and imprecise data.

In addition to data management issues of data streams, many other problems arise. Scientific measurements have standard errors. For example, location data for moving objects involves uncertainty in current position. Individual sensors are not reliable and, consequently, wireless communication is also unreliable. Thus, various approaches are used to provide more accurate estimation of the environment. In multisensor data fusion approaches like fuzzy sets or Dempster-Shafer evidential theory are sometimes used (Ramamrithan, et al., 2004).

Traditional DBMSs were applied to business data processing, which typically focused on numbers and character strings. In those application areas, data elements are precise quantities like address, quantity on hand, balance, status, and delivery date. As a result, current DBMSs have no facilities for either approximate data or imprecise queries. Sequences and images require approximate processing based on a similarity, metrics, etc.

Last but not least, to increase data quality new information processing occurs that preserves and retrieves the origins and processing history—that is, the *lineage*—of objects and processes (Bose/Frew 2005). To ensure that the greatest use is made of environmental data, data producers should include data lineage (and authenticity information) in the metadata. On a database level, this requires more sophisticated techniques for metadata processing.

5. Data mining and OLAP

Environmental data often need to be analysed in order to obtain information necessary for environmental management decisions. In comparison to simple forms of regularities/dependencies treated by statistical methods, data mining methods can find more complex hypotheses that include both numerical and logical conditions.

Historically, data mining has focused on efficient ways to discover models of existing data sets. These models must expose some useful aspect of the data, while obscuring details not useful for the intended application. Algorithms have been developed by many research communities to perform such operations as classification, clustering, association-rule discovery, and summarization. These techniques are now part of mainstream products from the major DBMS vendors and most of them are applicable in EMSs.

Often OLAP techniques are sufficient. For example, temperature and pressure trends are required in an environment. Derivation of such information typically requires past temperatures and pressures stored in a database and processed along the time dimension.

Recent interests in combining data mining technology with DBMSs require new approaches to storage datasets to be mined and to optimize data mining processing. New research directions include (1) multi-dimensional OLAP for discovering unusual patterns in stream data; (2) mining clusters and outliers in stream data for discovering unusual patterns; and (3) single-pass classification methods for stream data mining.

6. Wireless broadcast and mobile computing

Data broadcast is an attractive alternative to on demand access because it can broadcast data simultaneously to a large number of clients at a fixed cost. It is suitable for location-based services, which exhibit strong temporal and spatial locality in that clients within the neighbourhood and a certain time period tend to seek the same kind of information (Zheng/Lee 2005).

Since environmental data requires often to be disseminating timely to the user anytime and anywhere, a mobile environment is of increasing importance in this context. Particularly in periodic broadcast, data is broadcast periodically on a wireless channel. A mobile client listens to the broadcast channel and downloads the desired data from the channel according to a query issued from the user or a stored profile of interest on the client. Of course, these networks should be able respond to aperiodic queries.

We observe that location becomes a very important property of data and introduces a new dimension to data access methods. Traditional data access methods are not suitable for such computing and new researches redefine some well-known techniques, e.g. spatial queries, in the mobile environment with a particular emphasis on broadcast data.

7. Conclusions

Environmental data management, analysis, and communication are essential components of environmental characterization and decision making. DBMSs, the Internet, and associated web technologies have become an integrating force for these components.

In fact, today's DBMSs provide an universal architecture applicable to a lot of various types of tasks. By words of (Stonebraker 2005) "*one size fits all*". According to (Selinger 2005), data research challenges for the next decade include apart the other things:

- re-examine DBMS architecture and invent ways to scale more, without sacrificing user-visible availability or performance,
- learn what managing content is all about, what is needed and create new models
- treat metadata as a first class research**.**

In new architectures of DBMSs, separate engines rather "*made to measure*" are supposed according to requirements of various applications. Besides rather traditional applications – OLAP, data warehouses, and text retrieval, some candidates for a separate engine are:

- stream processing,
- sensor networks,
- scientific data bases,
- native XML databases.

We have tried to highlight some characteristics of the first three technologies with respect to their association to environmental data management. Everything indicates that the development of these technologies has and will have consequences which affect EMSs of the future.

Acnowledgement

This research was supported in part by the National programme of research (Information society project 1ET100300419).

Bibliography

Abiteboul, S., et. al. (2005): The Lowell Database Research self-assessment, in: Communications of the ACM, May 2005/Vol. 48, No. 5, pp. 111-118.

Amato, G., Caruso, A., Chessa, S., Masi V., and Urpi, A. (2004): State of the art and future directions in wireless sensor network's data management. 2004-TR-16, published by ISTI.

Bose, R. and Frew, J. (2005): Lineage Retrieval for Scientific Data Processing: A Survey. ACM Computing Surveys, Vol. 37, No. 1, pp. 1-28.

Bright, L., Maier, D. (2005): Deriving and Managing Data Products in an Environmental Observation and Forecasting System, in: Proc. of Conference on Innovative Data Systems Research (CIDR), January 2005, pp. 162-173.

Carney, D., et al (2002): Monitoring streams - a new class of data management applications, in: Proc. of VLDB, pp. 215-226.

Ramamritham, K., Son, S.H., Dipippo, L.C. (2004): Real-Time Databases and Data Services. Real-Time Systems, Kluwer AP, 28, pp. 179-215.

Selinger, P. (2005): Five Data Challenges for the Next Decade, key note of the Conference ICDE, held in April 2005, Tokyo, Japan.

Stonebraker, M. (1994): Sequoia 2000-a reflection on the first three years. Sequoia Technical Report S2K-94-58. Berkeley, CA. Available at: http://epoch.cs.berkeley.edu:8000/sequoia/techreports/ s2k-93-23/.

Stonebraker, M. (2005): "One Size Fits All" An Idea Whose Time Has Come and Gone, key note of the Conference ICDE, held in April 2005, Tokyo, Japan.

StreamBase Systems, Inc. (2005): StreamBase™ 2.0, available at: http://www.streambase.com/index.html

Tomasic, A. and Simon, E. (1997): Improving Access to Environmental Data using Context Information. ACM SIGMOD Record, Volume 26, Issue 1, pp: 11 – 15.

Zheng, B., Lee, D.L. (2005): Information Dissemination via Wireless Broadcast, in: Communications of the ACM, May 2005/Vol. 48, No. 5, pp. 105-110.

LCA (2005): Glossary. Available at: www.lineadecreditoambiental.org/html/glossary.html