

ESTABLISHMENT OF WATER ENVIRONMENTAL DATA MART

Z. L. Liao and Z. X. Xu

State Key Laboratory of Pollution Control and Resource Reuse, Tongji University 1239 Siping Road, Shanghai 200092, China, email: zl_liao@mail.tongji.edu.cn

ABSTRACT

The concepts of Data Warehouse, Data Mart, OnLine Analytical Processing were introduced into the domain of Environmental Decision Support System. A Water Environmental Data Mart (WEDM) was developed with a city's water environmental operation database as data source, and with star schema as data framework. Two main parts were included in WEDM, the Extract-Cleanse-Transform-Load tool, and the universal OnLine Analytical Processing tool whose functions include SQL, Classifying Statistic, and Data Visualization. The WEDM could provide multi-dimensional, multi-leveling, integrated, dynamic, flexible querying and analyzing, which was not offered by previous environmental operation database system. Two examples showed the functions of proposing decision support information of WEDM.

Key words: Water environment; Data Warehouse; Data Mart; On-Line Analytical Processing

1. INTRODUCTION

Information on pollution sources, hydrology and water quality monitoring, natural and social conditions in water environmental protection are available extensively. In recent years, water environment information management systems have been established and the automatism of daily transaction processing has been realized. As a result, water environmental protection decisions have been made in some extent by simple data querying, making statistics and computing and using visualized tools such as GIS (Zhang, *et al.*, 2005).

Nevertheless, these water environmental management information systems are mainly employed for routine transaction processing, and only single database, details, presence and speed are concerned. Whereas for the decision makers in the executive level, macroscopically and integrated information and historical influences are more concerned. The process of decision analysis usually involves more than one data source. Furthermore, dynamic data integration and multi-leveled synthesis are needed, historical data should be analyzed and many system resources may be occupied in such a process (Zhang, 2004).

In recent years, with the prevalence of database's application, it has been gradually realized that analytical processing is different from transaction processing and problems will appear when transaction processing system is used to directly support decisional analysis. In this case, data warehouse technology appears.

Data warehouse is a new technology to organize, save and access data by multidimensional and integrated methods. It can pool and process data from different sources to form a uniform data source. Its characteristics are theme-faced, integrated and stable. End-User of data warehouse can make multidimensional querying and analyzing, and can realize data information visualization (Mallach, 2000; Sid, *et al.*, 2003).

The concept of data warehouse was proposed in the early 1990s. By the mid of 1990s, it became the international fad. Recently, it is widely used in commerce, finance and the telecom industry. However, there are few reports in the world of data warehouse in the environment protection field (Vassiliadis, *et al.*, 2001). In contrast, with data warehouse high cost and wide work scope, data mart is a smaller and more centralized kind of data warehouse and is more frequently used. It is a set of data and operation rules organized for specific user's decision support (Chen, *et al.*, 2002).

The data warehouse/data mart can be developed and applied through extracting, cleansing, transforming the data from multi-source (databases) according to the designed data warehouse structure, and loading into data warehouse. Existing or newly developed On-Line Analytical Processing tools and data visualization tools, as well as data mining tools to query, make statistics, display and mine relevant knowledge for the purpose of getting corresponding decision support information, are thus established and custom-applied.

The objective of this work is to develop a data mart named WEDM (Water Environment Data Mart) based on the operation database of Shanghai urban water environment. Two tools, i.e. Extract-Cleanse-Transform-Load tool, and universal On-Line Analytical Processing tool are developed. The functions of WEDM, i.e. multidimensional, multi-levelled, integrated, dynamic, and flexible query and analysis, are displayed by two examples.

2. MODEL DESIGNING OF WATER ENVIRONMENTAL DATA MART

Data warehouse is organized in multidimensional structure. Most of the data warehouses are using the so-called “star schema model”. This model is centered on a fact table containing measures with related dimension tables which characterize these facts. Each dimension has a number of attributes used for selection or grouping. It is found that the star schema is very fit to data mart.

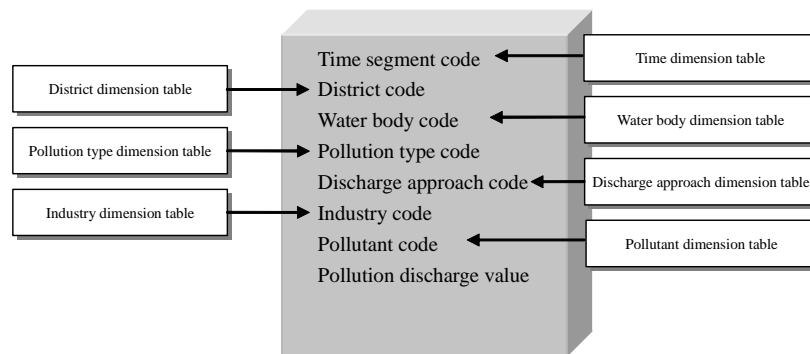


Figure 1. The pollution source star schema developed

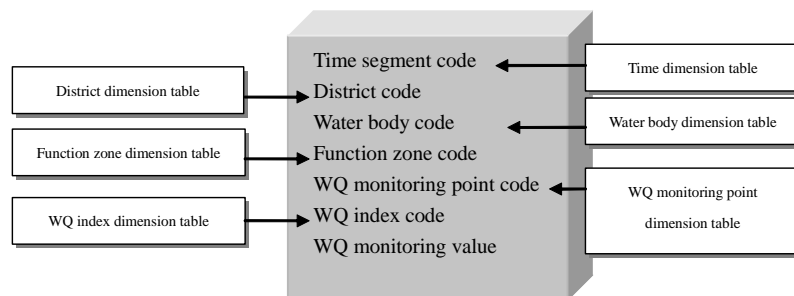


Figure 2. Water quality monitoring star schema developed

In WEDM, star schema is used to organize data. According to the situation of operation database of Shanghai urban water environmental and the requirements of water environment rehabilitation, the

following three fact tables are determined: pollution source, water environment (including water quality monitoring, hydrology monitoring, water body of river, etc.) and rehabilitation project. The data tables of star schema are linked to each other with the same dimensions, such as time dimension, river (water body) dimension, etc., and formed a “star-net query model”.

By linking operation on fact table and several dimension tables with the main code of dimension tables, the data value and its multi-dimensional descriptions can be obtained in one query.

As a primary exploration, only pollution source data and water quality monitoring data are included in WEDM. Their star schemas are shown in Figures 1 and 2.

3. ECTL PROCESSES AND TOOLS

After the data mart model is designed, the data structure is determined accordingly. The data of Data Mart usually comes from an existing operation database (often more than one source). Nevertheless, there are differences in data structures, data integrity and consistency between database and data mart. The ECTL processes must be done before loading the data from the database into data mart.

ECTL are the processes of Extracting, Cleansing, Transforming and Loading data from the source database according to data mart’s structure. After ECTL, objective data source, i.e. the data mart, can be obtained.

Traditional Shanghai urban water environment database has problems such as unclear field meanings, non-custom field type and null or zero values. Relevant data tables are chosen, including “water environment pollution sources in past years,” “water quality monitoring stations,” and “water quality monitoring data in past years.” The “water environment pollution sources in past years” is transformed to WEDM’s “pollution sources fact.” The “water quality monitoring stations” and the “water quality monitoring data in past years” are merged and transformed to WEDM’s “water quality monitoring fact table.”

There is no table about rivers, reaches and rehabilitation projects in the operation database of Shanghai urban water environment. Three null tables are established for the purpose of the expansibility of the data mart. They are the table of river, the table of river reach, and the table of the rehabilitation project,

There are a lot of difficulties if only depending on manual work to implement the processes of ECTL on the data of source databases. It will waste a lot of time for the magnitude of data tables. For example, there are 110,000 records in the water pollution sources database, and the record amount will increase continuously. On the other hand, with the low efficiency of manual work, it is easy to make mistakes. Therefore, a data transform tool of water environment database is developed to solve the data transforming issue of source data.

By this transformed tool, data is extracted, cleansed, transformed and loaded into data mart. Hundreds of repeated or wrong records are eliminated.

4. DEVELOPING OLAP TOOL

After establishing Data Mart, how to operate it and obtain valuable information from it becomes the concern of a decision maker. The traditional database tool is OLTP—On-Line Transaction Processing. But it is directed to operator and lower manager: the main operation is querying, adding, deleting and modifying the data of database to finish the transaction processing, and the characteristics should be quick response and frequent modifying. But data warehouse and data mart are directed to

decision maker and higher manager, who will distill all-sided, comprehensive and time-varied information to support decisions. Therefore, it is insufficient to only depend on OLTP. OLAP—On-Line Analytical Processing must be used.

OLAP is a new query technology. The analyzer can flexibly, quickly, consistently and mutually observe information from multidimensional angles for the purpose of understanding data. Its basic operation includes slicing, dicing, pivoting, drilling, data visualization, etc. The technology kernel of OLAP is the concept of dimension. Therefore OLAP can be regarded as the collection of multidimensional data analyzing tools.

A data analyzing software is developed to conduct OLAP analysis in WEDM. The functions of OLAP of WEDM involve three parts: SQL query, classifying statistics and data visualization.

(1) SQL query

Query function is one of the basic functions of data analysis. This function can execute all sorts of condition querying, including numerical value, character, field, sort, delete duplicate data, restrict record number, etc. Query function is established on the basis of Structured Query Language (SQL).

(2) Classifying statistics

Water environment data includes all kinds of meaningful numerical fields and multidimensional description of these fields, such as time, terrain, direction, type, etc. Classifying statistics from multidimensional angles is necessary in decision analyzing. Specifying the classifying fields, a user can make integrated statistics on numerical fields, such as summary, averaging, computing the square error, etc.

(3) Data visualization

Data visualization is an important part of OLAP. A data visualization tool is developed in WEDM. Using all kinds of data from data mart, histogram and scattering chart of relevant fields can be described with visualization module, and the data can be exhibited to user intuitively.

5. WATER ENVIRONMENTAL DATA MART APPLICATION EXAMPLES

(1) Example one: For Shanghai City, year 2000 and 2002, the discharging approach is “direct discharging into river.” Make statistics of the total amount of sewage per year. The classifying statistics are according to pollution sources.

Open the pollution source table, and operate it according to the following steps.

Step 1: Input querying condition: “year=2000, 2002; discharging approach=direct”, then all conformed data will appear in data analyzing windows.

Step 2: Make classifying statistics according to pollution type.

Step 3: Choose “total amount of wastewater per year”, and “depict histogram”.

The results are shown in Figures 3 and 4.

Thus, user can compare total amount of wastewater each year of all kinds of pollution sources of direct discharging into river in year 2000 with in year 2002. Some useful information can be obtained. For example, the amount of livestock and poultry wastewater directly discharged into river in 2002 reduced 73% in contrast with that in 2000, but the amount of industrial and residents domestic wastewater increased in some extent, and the amount of enterprises domestic and livestock & poultry wastewater decreased a little.

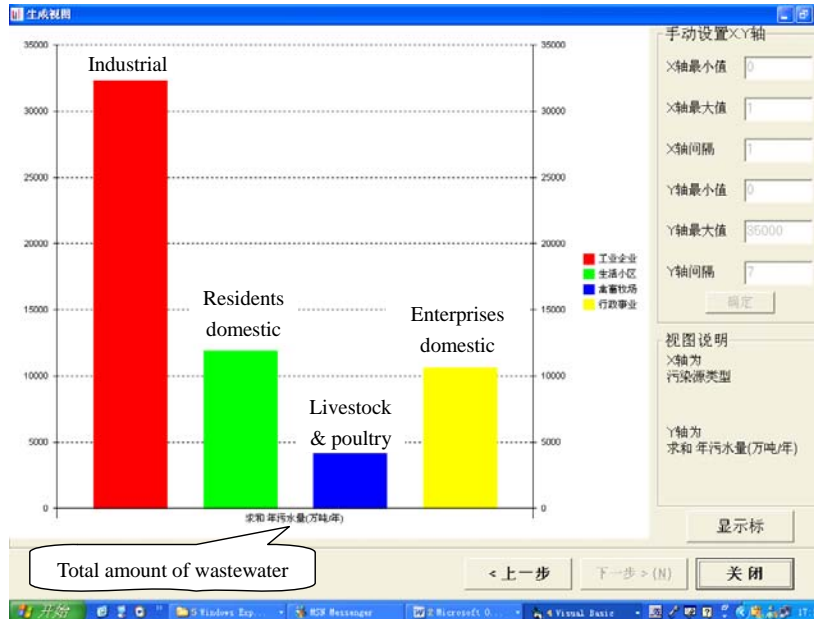


Figure 3. Total amount of direct discharging wastewater of all kinds of pollution in year 2000

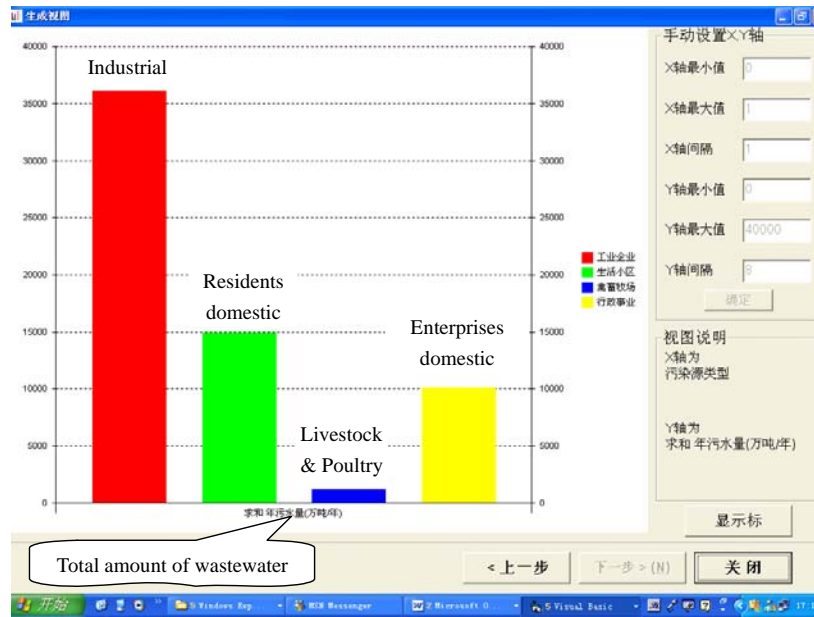


Figure 4. Total amount of direct discharging wastewater of all kinds of pollution in year 2002

(2) Example two: Examine the average permanganate index value of all monitoring stations in the city from 1995 to 2002.

Open the water quality monitoring table. Input the following querying and statistics conditions: “from year 1995 to 2002” and “the average value of water quality monitoring in the whole city”. Do visualization operation on the results and choose “histogram”. The results are shown in Figure 5.

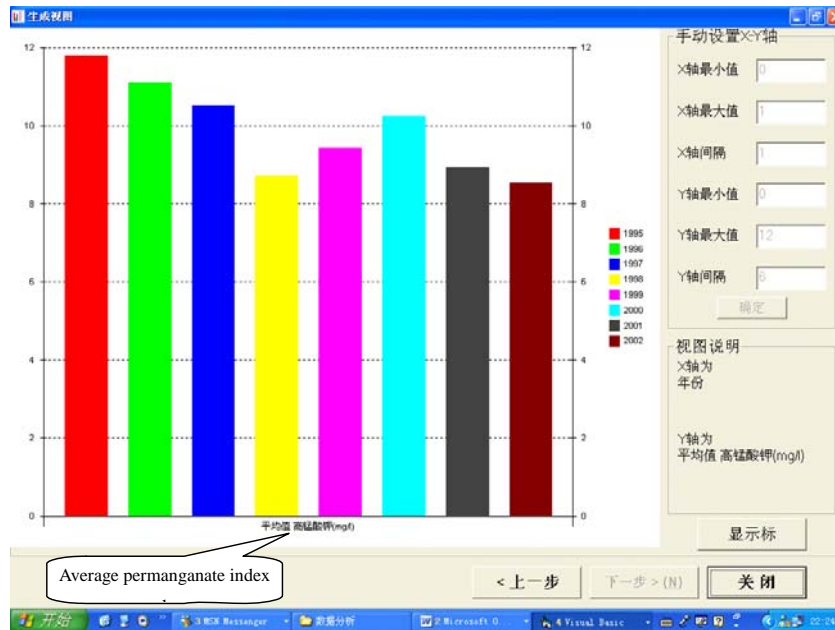


Figure 5. The whole city average permanganate index value from year 1995 to 2002

From Figure 5, it can be found that from 1995 to 2002, the whole city average permanganate index value appeared descending trend as a whole. It illuminates that the water environment became better.

From the above two examples, we can see that WEDM has basic functions of OLAP, and can distill useful information from messy data to support decision making. User can specify querying condition at will, and make classifying statistic and visualization on the querying results. For unsatisfied querying results or better querying ideas, user can do secondary querying. For satisfactory querying results, user can save them. Moreover, the query, manual work statistics and visualization modules of the software can commonly be used on any database in ACCESS format. These will be the basis of common usage of OLAP tool for water environmental data warehouse in the future.

6. SUMMARY AND PROSPECTS

Not only the necessities of establishing water environmental data warehouse are set forth, but also a probing is done. Based on the operation database of Shanghai urban water environment, a Water Environment Data Mart including ECTL and OLAP tools is developed. With these tools, multidimensional, multi-leveled, integrated, dynamic, and flexible query and analysis can be done.

In environment field, Management Information System (MIS) has been used widely, but data warehouse, OLAP and data mining are used little. With the recognition of environment protection in China day by day, the requirements of environment decision level will become higher and higher. On the other hand, the information of environment data is increasing quickly. Therefore, importing new technologies of MIS and Decision Support System (DSS) such as data warehouse/data mart must become more popular.

Data mining is not involved in this paper. Data mining can be seen as the higher phase of OLAP and it is the process of discovering or mining deeply implicit information or knowledge. It has many successful cases in commerce domain. The most famous one of them is the “beer and diaper” story from Walmart. Data mining is also an indispensable tool for environment database’s development and application in the future.

REFERENCES

- Chen Y.X., Bao H. (2002). Data Mart Technology. *Micro Computer Development*, 6: 23-25. (in Chinese)
- Mallach, E. G. (2000). *Decision Support and Data Warehouse Systems*. McGraw-Hill Companies, New York.
- Sid A., Larissa T.M. (2003). *Data Warehouse Project Management*. Tsinghua University Press, Beijing. (in Chinese)
- Vassiliadis P., Quix C., Vassiliou Y., et al. (2001). Data Warehouse Process Management. *Information Systems*, 25(3): 205-236.
- Zhang Q.Y., Tian W.L. and Shen X. (2005): *Environment Management Information System*. Chemical Industry Press, Beijing. (in Chinese)
- Zhang Y.F. (2004). *Decision Support System*. Wuhan University Press, Wuhan. (in Chinese)