DATA WAREHOUSE AND TRANSACTION INFORMATION SYSTEMS

Bohuslav Martiško

Fakulta financií, Univerzita Mateja Bela, Cesta na amfiteáter 1, 974 01 Banská Bystrica, SR
bohuslav.martisko@umb.sk

Abstract

The paper deals with the relation of classic (primary) business information systems (ERP) with management systems (MIS, EIS). It concerns the data warehouses building that are the source of input data for management systems. The difference between the content and the organisation of relational database is briefly outlined. The database is utilised by primary information system and the data warehouse where besides relational databases mainly post-relational and multidimensional models are used.

Key Words: IS - Information System, OLTP, OLAP, EIS, Relation Database System, Multidimensional Database System, DATA WAREHOUSE, DATA MART, DATA MINING, Bitmap Indexes, CACHÉ

1. Characteristics of a DATA WAREHOUSE

This part of information systems is becoming more important. Requirements on the work of managers are always higher. They take important decisions, which influence the economic activity of firms. Therefore, managers need high-quality bases to be able to make such decisions.

A data warehouse is a subject-oriented informational database designed specifically for reporting, analysis and decision making. Here's how a data warehouse differs from transaction systems (OLTP) :

OPERATIONAL DATA **OLTP –TRANSACTION SYSTEMS**	**BUSINESS INFORMATION** **OLAP - DATA WAREHOUSES**
Transaction (application)-oriented	Subject-oriented
Hold detailed data	Summarized data (+ detailed)
Only current data	Historical data (+ current)
Data continuously changing	Data stable over time
Many concurrent users	Few concurrent users
Simple queries and updates	Complex queries
Need rapid response times	Can tolerate longer response times

There are used the following methods of analyses:
- drill down - decomposition from top downwards
- drill up - building from sub to top
- drill across - transfer from one dimension to another.

2. Transaction Systems (OLTP) and Data Warehouses (OLAP)

In general, we can define three levels of management:
- operational - dispatcher, solves operational tasks
- tactical - level of department managers
- strategic - level of top management. Main decision making about the firm or organization.

Strategic and tactical management is very important, at present, especially in middle and large firms, organizations and financial institutions. New technologies of data warehousing, EIS, MIS and DSS systems [3] enter into this field. Basic (primary, transactional) information systems solve the problem of evidence and provide state legislative with relation to enterprises. They create a wide database, sometimes very heterogeneous (databases of different producers or data files), which includes mostly primary data. This database is used mainly for standard evidence and operative management. It is not suitable as a direct impute into the EIS systems, used by middle and top management. Creation of high-quality bases for manager decision-making directly depends on the quality of inputs. Therefore, there is created a new database (secondary), which fulfills the requirements of creating a high-quality bases for management by content and also by organization. This kind of database is called Data Warehouse (DW), The activity of creating such database is called Data Warehousing. If data warehouse contains information from the field of interest of the certain department, it is called Data Mart (DM) - a sub-data-warehouse, data market.

It is not only a mechanical data transport from the primary data sources into the data warehouse. The data are always also processed, so that it increases their informational value. The data warehouse consists not only from the primary data (there may be some exemptions coming out of detailed views), but also from derived data, aggregations, assumptions, time series, etc. It is very important to propose a correct content of DW (even if we admit the problems connected with physical transformation of data structures, what is solved by computer experts). It requires experienced and skilled workers especially in the field of economics (of economic analysis). They set the contents of DW. The whole work would be just a waste of time if analysis were incorrect, even if the technology of data-transformation process was carried out excellently by computer experts, and it would not satisfy the requirements of managers. The presence of computer experts concerning certain EIS tools is required only during installation and the first setting of the system. Finally, it is connected with a trend of automation of programming. New development systems cover a large part of routine activities of "classical" program-creators, and workers with education in economics manage part of the work (of application software development) along with their own tasks.

A standard relation database system may be also used as database for the data warehouse. In the mean time, special multidimensional organization of databases is used for this purpose very often. There are solved problems with of more dimensions, mathematical problematic of sparse matrices, and the use of different compressed algorithms. We know two alternatives in creation of multidimensional databases:

- common hypercube, in which all dimensions are stored. Volume is increasing exponentially by the increase of the number of dimensions. It is non-economical, and therefore it is used for smaller applications.
- more data hypercubes. This form is used more often.

Indexes (bitwise or bitmap) and summarization occupy 75% of DW. The rest consists of base-data.

The data are enriched sometimes by more dimensions by transport from primary sources. These dimensions are added on the base of requirements of managers. Database fields may include long character strings, by means of which the size of database increases (until terabytes). This fact causes besides disk space problems (these must be not significant, today) also long access time during scanning databases. This problem is solved differently by many different firms. For example, by the use of bitmap indexes. Principle is as follows:

2.1 Relation database

number of record	sex	insurance	county,...other fields	
1	women	yes	Zvolen	.....
2	women	no	Zvolen	.....
3	man	yes	B.Bystrica	.....
4 until10M rows	women	no	Zvolen	....

length of 1 row = 800 byte

By means of logical evaluation of individual rows, bitmap fields are created during the indexing process, which runs as batch in the free computer capacity (for example during night).

2.2 Bitmap Indexes

number of record	sex	insurance	county
1	1	0	1
2	1	1	1
3	0	0	0
4 to 10M bitov	1	1	1

length of scanned row = 3 bits

If we have a question for example "How many women are unemployed in Zvolen city?", we must scan in relation databases a large amount of data (sometimes the whole table scan passes through). It takes a long time, because there are many accesses to the disk and many I/O operations. In our case the calculation is as follows:

800 bytes x 10M rows / 16K blocks = 500 I/Os

During DW scanning (by bitmap indexing):

(10M bits/8) x 3 columns / 16K blocks = 240 I/Os

The answer to our question is "2 women" (highlighted rows 2 and 4).

This way is suitable for pre-defined queries. These involves around 51% of all queries. It means, that the rest are ad-hoc queries. Answering these queries lasts longer time. It is suggested to re-calculate the indices, if some ad-hoc query is often repeated.

3. Post-relational database CACHÉ

Caché is a post-relational database that uniquely offers three integrated data access options:

- a robust object database,
- high performance SQL,
- and rich multidimensional access.

This database system enables rapid Web application development, extraordinary transaction processing speed, massive scalability, and real-time queries against transactional data - with minimal maintenance requirements.

Caché is available for Windows, OpenVMS, Linux and major UNIX platforms and it is deployed on systems ranging from two to over 10,000 simultaneous users.
It contains many new and improved features. Here is a brief description of a few of the most exciting changes.

Caché Studio
Caché's user interface has been revamped to make it even easier to develop applications quickly. The new Caché Studio is an integrated development environment (IDE) that allows developers to create objects and Caché Server Pages, write and debug methods and routines, expose data as Java classes, XML documents, etc.

Integrated XML Support
Caché is an excellent match for XML, which is quickly becoming the new standard for sharing data between applications. Caché 5 comes with system level integration that enables bi-directional compatibility with XML. XML documents and DTD files or XML schemas can be created from Caché classes, and Caché objects can be created from incoming XML documents. This all happens automatically, without the need for developer-defined mapping of XML to internal database structures.

Web Services and SOAP
In addition to sharing data, many enterprises want to be able to share functionality by publishing (and consuming) Web Services. Now, Caché methods can easily be published as Web Services. Caché will automatically generate a WSDL descriptor for the service, and when the service is invoked, Caché will send the response as XML, formatted according to the SOAP protocol.

Enterprise Java Support
Caché has long included a Java binding for Caché classes, but now, Caché classes can also be exposed as Enterprise Java Beans. Caché's EJB binding makes it easy for developers to use Bean-Managed Persistence, eliminating the need for mapping objects to relational tables.
Transactional Bit-Map Indexing

It is the first database to enable real-time data analytics by supporting bit-map indexing for "live" data. Bit-map indexes are known to significantly improve complex query response, but updating them has traditionally been so slow that they have only been used in data warehousing applications. Caché's new bit-map indexing capability is so nimble, it can be used even with rapidly changing transactional data.

Conclusion

Problematic of Data Warehousing is a very wide area. It intervenes into many different areas. It is necessary to manage the form of drawing data from heterogeneous data-sources (different databases, operation systems, data files) to. Economic sciences must guarantee a correct definition of analyses, what defines the meaning of the whole activity. Mathematics is applied during definition of DW structures, building of indexes, effective comprimation-methods and designing.

Lately, the term Data Mining appears in context of DW building. It seeks for hidden connections, relations and dependabilities in data. There applied methods of artificial intelligence and neural networks.

Literature:

1. Martiško, B.: Od transakčných systémov k manažérskym aplikáciám, CSSI-CVIS, marec 2003, Univerzita Tomáše Bati ve Zlíně, str. 126-131
2. Imon, B.: WEB-oriented Data Warehousing ISBN1-02016-390-7, Cadenhead Publishing, 2001
3. Lacko, B.: Aplikace progresivních technologií a technik pro počítačovou podporu rozhodování, zborník z medzinárodnej konferencie Svět informačních systému, UTB Zlín, marec 2004

Poznámka vydavatele:
Příspěvek neprošel gramatickou ani terminologickou redakcí v anglickém jazyce.