Database Vs Data Warehouse

Manole VELICANU, Bucharest, Romania, mvelicanu@yahoo.com
Gheorghe MATEI, Bucharest, Romania, george.matei@bcr.ro

Data warehouse technology includes a set of concepts and methods that offer the users useful information for decision making. The necessity to build a data warehouse arises from the necessity to improve the quality of information in the organization. The date proceeding from different sources, having a variety of forms – both structured and unstructured, are filtered according to business rules and are integrated in a single large data collection. Using informatics solutions, managers have understood that data stored in operational systems – including databases, are an informational gold mine that must be exploited. Data warehouses have been developed to answer the increasing demands for complex analysis, which could not be properly achieved with operational databases. The present paper emphasizes some of the criteria that information application developers can use in order to choose between a database solution or a data warehouse one.

Keywords: *data warehouse, database, database management systems, information systems, data organisation in externe memory, business intelligence.*

The need for databases and datawarehouses

The possibility for users to efficiently access data through analytical queries is of extreme importance for the competitive advantage of companies. The transfer and sharing of data within the organisation, among departments and different locations as well as among business partners is also important. Solutions are more and more numerous due to the multitude of systems which may be integrated with decision support systems: databases, data warehouses, data marts, business intelligence solutions, enterprise-based applications. Managers who will succeed are those who will implement decision support systems, business performance monitoring applications, executive information systems or business intelligence solutions [INMO05a]. Through these technologies, companies learn what has happened to their business, why has happened and what will happen; and all this, plus the expertise and intuition of users generate competitive advantages.

Organisations must store and process more and more data, which is more and more diversified. The need to use this data as a resource for the organisation, as decision support, has triggered the ongoing improvement of information systems. The better the data of a company is organised, the better the company results.

The organisation of large volumes of data has evolved from files to database and latter to data warehouses. The pre-requisite of storing and processing larger and larger volumes of data has led to the design of analytical systems based on data warehouses.

The purpose of such a system is to provide analysts with an integrated and consistent view on all the data relevant for the company. Based on the data systematised and consolidated in such data warehouses, comprehensive analyses of the performance of the company may be derived, various data correlations may be identified, together with trends which forecast future developments, as well as solutions for the improvement of business.

The increased need to anticipate changes in market conditions and customer choices requires the development of intelligent business plans, which involves access to necessary information. Most of this information can be found in transactional systems, relational databases included. The ability to transform data into information, information into knowledge and knowledge into action [VELU03] is a must, so that companies may be competitive in an ever-changing economic environment. The solution to all such problems is the development of a data warehouse.

Transactional databases provide an answer to *operational requirements,* while data warehouses provide an answer to *analysis requirements,* thus offering the possibility for high quality analyses and complex ad-hoc queries through user friendly interfaces.
The fundamental criterion for the organisation of data in data warehouses is the subject (the field of activity), while the fundamental criterion for databases is the application.
The data warehouse concept is a *logical architectural approach* to extracting operational data and transforming it into accurate historical information to support the decision-making process.

Comparative view of the features of data warehouses and databases

The features of two data organization mode in external memory, data warehouses and databases, can be seen from the definitions provided below.
A *database* is an application-oriented collection of data that is organised, structured, coherent, with minimum and controlled redundancy, which may be accessed by several users in due time [VELU03].
A *data warehouse* is a subject-oriented collection of data that is integrated, time-variant, non-volatile, which may be used to support the decision-making process [INMO05b].
We may thus infer the main features of data warehouses and databases, which shall be described below.

Subject-orientation

Data organisation in data warehouses is based on areas of interest, on the major *subjects* of the organisation: customers, products, activities etc. Databases organise data based on enterprise applications resulted from its functions.
The main *objective* of a data warehouse is to support the decision-making system, focusing on the subjects of the organisation. Information is organised on these subjects. The objective of a database is to support the operational system and information is organised on applications and processes. All data items which refer to the same subject or event in the real world are linked and the orientation - data to process - is obvious in the content of the database.
The data warehouse includes *only information* which shall be employed in information and analysis processing, while the operational database includes the detailed data required for processing purposes, but which is not relevant for management or analysis.
The subject-orientation of a data warehouse allows the development of a decision-making process through an *incremental process*, which integrates different subjects into a single structure. For example, when a customer is included in several operational databases, where he is differently defined, the customer is defined only once within the data warehouse and is viewed by all users in the same way.

Integration

Operational databases are designed at different times, by different teams, in different ways. Thus, from a functional point of view, the database cannot be employed for analysis and reporting purposes. Data warehouse is an enterprise project. It includes data from all or most operational databases of the organisation, which is stored consistent to allow analysts to focus on the use of data, not on its reliability and consistency. *Consistency of data* is very important for databases and is ensured by the objective of the database management system - DBMS regarding data integrity. Consistency also applies to data warehouses with respect to: field names, code systems, date representations, variable measurements, physical attributes etc., so that the reports generated for various departments or different times shall include the same results.

Time-variant

The value of operational data in databases is updated periodically and shows the current status. On the other hand, for the information requirements of economic analysis based on data warehouses, historical data is of the essence, as it shows the trends for accurate forecasting. The regular loading of data from operational databases makes the data in data warehouses time-variant. Data in data warehouses accurately shows the status at different moments, thus providing a historical view

of date. This makes data warehouses different from operational databases, where data should show the status at the time of access.
In databases, data is *updated* with each new transaction and former values are usually lost. Operational databases rarely keep historical data and this only for short times, as their purpose is to keep current data. As opposed to these systems, data warehouses are not updated, but data is loaded periodically to show the history of data. This allows for the identification of trends, as well as for comparisons between different time periods. The time horizon of data warehouses is significantly longer compared to that of operational databases, providing information from a historical perspective (5-10 years). That is why, any structure in the data warehouse includes, either explicitly or implicitly, the time element, to identify a certain feature at a certain time, which is not mandatory for databases.

No volatility

Data in data warehouses is static, not dynamic as is the case with operational systems. As data warehouses show operational data at a certain time, data will not be updated once loaded in data warehouses. As a result, an identical query made after one year based on the same reference data will yield the same result. In operational databases, information is volatile, as queries focus on current data. Data is updated on an ongoing basis, usually on a transactional basis. Any transaction processed involves updating: adding new records, modifying or deleting existing other.

Differences between data warehouses and operational databases

Operational databases and data warehouses are mostly based on the *same technological support*: both are data collections, both function based on keys, indexes and views, both is based to a data model.
Nevertheless, the two systems *are different*, as the criteria described below shows.
1) From a *functional* point of view: operational databases process transactions, providing answers to operational requirements, while data warehouses are used based on ad-hoc queries, mainly for management purposes.
2) *Functional requirements* are different: operational databases mainly focus on data security and coherence, which makes queries slow, special ad-hoc, mainly in the case of unpredicted criteria, while in data warehouses is usually. These queries, specific to economic analysis, may significantly compromise the performance of the operational system, due to the lack of predictable indexes, as is the case of data warehouses.
3) Although most operational systems and data warehouses are built on relational technologies, their *design* is substantially different, as their purpose is also different. Operational databases are designed for online transaction processing and their main objective refers to the efficient storing of a large amount of transactional data. They include current information on day-to-day activities and process-oriented information which is subject to updating. As a result, data is dynamic and thus, very volatile. The tasks of such systems are structured and repetitive and are made up of current, short and isolated transactions, which include detailed data. These transactions read or update few recordings – tens at most, mainly accessed based on their primary keys. Operational databases reach sizes from hundreds of megabytes to gigabytes. Their consistency is essential and refers to rapid transaction processing.
As opposed to transactional databases, data warehouses are designed to be the support of decision-making systems. They are designed to facilitate data analysis, not efficient storing, and the only operations performed refer to the initial data loading, data access and its refreshment. Historical, summarised and consolidated data is more important than detailed data. Data warehouses include consolidated data from several operational systems, with long time horizons, and have an integrated and evolutive vision. As data is static and non-volatile, the size of data warehouses may reach in time hundreds of gigabytes, terabytes or even petabytes. Many ad-hoc queries may be made and millions of records

may be accessed, as many joins and aggregations may be performed. Information is subject-oriented, as data warehouses provide a multidimensional view on data, based on an intuitive model, designed to meet the requirements of data analysis and decision makers.

4) Another difference refers to the *status* shown by data. Data warehouses show the status of data at different moments in time, thus providing historical outlook. This is different from operational databases, where data shows the current status at the time of access.

5) *Optimisation* is an issue for both systems, but in a different way. While operational databases are designed to provide data processing optimisation and security, data warehouses optimise analyses and the economic significance of data. Operational databases can be said to be optimised for writing, while data warehouses are optimised for reading. To this purpose, multidimensional modelling is used for the design of data warehouses to make queries for the analysis and summary of large amounts of data more efficient. The structure of a data warehouse is simple, intuitive and easy to understand by non-expert users, as opposed to the structure of an operational database, which is designed based on the entity-relationship model, by specific techniques which are complex and difficult to understand. On the contrary, multidimensional modelling involves the denormalisation of tables. This enters controlled data redundancy, allows analyses from different points of view and different levels of detail.

6) *Categories of users* are different. Operational databases are meant for a large number of users, from different categories. Automated processes are repetitive, as processing requirements are known before the initial development. The system should immediately provide answers to any query or to any new transaction. As opposed to these systems, data warehouses are used by a small number of users, namely by managers and business analysts. Processes are heuristic, as requirements are not completely known before the initial development. Response requirements are more lax compared to operational systems. Depending on the complexity of processing requirements, response times ranging from several seconds to days are allowed.

7) *Data integrity* is seen differently. Integrity constraints are established for the verification of input data in operational databases. These constraints are not necessary in the case of data warehouses, as data has been verified and filtered before loading, and historical data will not be updated following its loading in the data warehouse. In operational databases, a transaction must lead the data collection from one consistent status to another consistent one. This involves complex mechanisms for data integrity maintenance systems: data logs, data restore, detection of blockings, backup and recovery. These mechanisms are useless in the case of data warehouses. Treating of updating anomalies are not as important as in the case of transactional systems, as data warehouses are specialised and optimised for the fast retrieval of large volumes of data, and updating refers only to the regular add of new data.

8) Another difference between the two types of systems refers to the mechanisms required for users' concurrent access. As data warehouses are not updated, transaction management, concurrent access management and other such mechanisms integrated in the database management system are used only in the initial loading stage and for subsequent add, due to the fact that they are expensive from the point of view of response time. These mechanisms may be disabled during the current use of data warehouses. The freedom thus generated may be employed for the optimisation of data access by: denormalisation, summarisation, data access statistics, index dynamic reorganisation etc.

9) *Backup and recovery* strategies are different for the two types of systems. Most data in data warehouses is historical data, which is non-variant and does not require repetitive saving. New data can be saved at the time of loading. It is advisable for data to be saved from intermediary databases in certain cases in order to minimise impact on the performance of data warehouses. Recovery policies may also be different in the case of data

warehouses as opposed to operational databases, depending on how critical permanent, seamless access to data warehouses is for the organisation. In actual database backup and recovery task is for DBMS. In actual data warehouse this task is for database administrator.

10) Not only data organisation is different in data warehouse from operational databases, but the interfaces used as well. Data warehouses support analytical processing by OLAP – On-Line Analytical Processing, which differs from a functional point of view from transactional processing applications. The attempt to perform analytical processing and comprehensive queries in operational databases will only reduce performance.

Conclusion. The differences shown above are part of the reasons why data warehouses are built separately from operational databases. The separation of the two systems ensures the scalability of business intelligence solutions as well as their ability to answer rapidly and efficiently to queries on the company. Data warehouses allow comprehensive analyses, as the structures of data collections: are more simple – only necessary information is retain, are standardised – structures are well documented, and are denormalised – there are fewer joins between data collections.

References

[1] [DEVL97] Barry Devlin, *Data Warehouse from Architecture to Implementation*, Addison-Wesley, 1997

[2] [ENWIKI] Enciclopedia Wikipedia, *Data Warehouse*, www.en.wikipedia.org

[3] [GAVA99] S. Gatziu, A. Vavouras, *Data Warehousing, Concepts andMechanisms*, Revista Informatik-Informatique, Nr. 1/1999.

[4] [INMO05a] Bill Inmon, *Data Warehouse and Decision Support Systems*, 2005, www.dssresoures.com

[5] [INMO05b] W. H. Inmon, *Building the Data Warehouse*, 4th Edition,Wiley Publishing, Inc., Indianapolis, 2005

[6] [KIMB02] Ralph Kimball, *The Data Warehouse Toolkit, Practical Techniques For Building Dimensional Data Warehouse*, 2nd Edition, John Wiley&Sons, New York, 2002

[7] [TUAR01] E. Turban, J. Aronson, Ting-Peng Liang, *Decision Support Systems and Intelligent Systems*, 7th Edition, Prentice Hall, Inc., 2004.

[8] [VELI05] M. Velicanu, *Dicţionar explicativ al sistemelor de baze de date*, ed. Economică, 2005.

[9] [VELU03] M. Velicanu, I. Lungu ş.a., *Sisteme de baze de date – teorie şipractică*, ed. Petrion, 2003.