CONSIDERATIONS REGARDING THE DATA WAREHOUSE, DATA MINING AND OLAP CONCEPTS

Adrian COJOCARIU, Cristina Ofelia STANCIU
„TIBISCUS" UNIVERSITY OF TIMIŞOARA, FACULTY OF ECONOMICS

Abstract:
Data warehouse is a complex system which contains operational and historical data concerning an organization, data provided by internal and external sources of the organization. The data warehouse overtakes data from the operational data base, the data being processed and analyzed in order to support the decision system. The main means to benefit by the data from the data warehouse are the on-line analytical processing (OLAP) solutions and Data Mining techniques.

Key words: *data, metadata, data warehouse, decision support systems*

JEL classification: D80, M15

At any level of an enterprise one has to deal with very large amount of data, provided by internal and external sources of the company. Internal sources are represented mainly by the manufacturing system of the company, while external data sources are represented by partners, clients, environment, market etc. The amount of data provided by internal sources is superior regarding to the amount of data provided by external sources, but the latter is increasing due to the development of some advanced techniques of data collecting.

The large amount of data of an organization must be safely stored, in order to be explored, and the main storing means are the data warehouse and data mart.

The data warehouse is a complex system that contains the operational and historical data of an organization, being separate from the other operational data bases. The enormous amount contained by a data warehouse comes from internal and external sources. The data warehouse overtakes the data from operational data bases, data on which different analysis will be made in order to support the decision maker within the decision process.

According to W. H. Inmon, the most important researcher in the data warehouse domain, they are "a collection of subject oriented, integrated, historical and persistent data, organized as a support for the decision process" [8], therefore the characteristics of data warehouse are: subject orientation, integration, historical character and data persistency. The subject orientation is a characteristic that results from the possibility of reorganizing data according to area of interest, and offers the advantage to develop decision support systems using an incremental approach. Each area is integrated in a unique structure, known as data mart. Data warehouse integration involves solving problems that result from the heterogeneous character of warehousing systems, and is a very complex, time and money consuming activity. The historical character of data warehouse derives from the need of storing all the values of the data along the time, because the evolution of data is important for the decision process. Data persistency is a

consequence of the historical character of data, considering that these data will only be queried, not modified too.

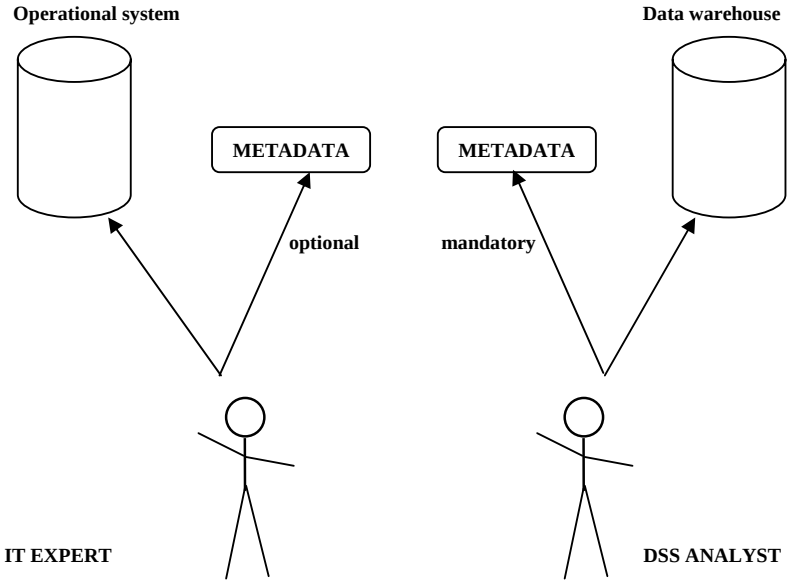

Figure 1. Using metadats within operational systems and data warehouse (Source: [7])

Using data warehouse proves to have a lot of advantages, worth mentioning being the following:

- decision makers easily obtain a series of reports that support the decision process;
- the data consistency, their productivity is increasing and the computational costs are decreasing;
- users may access a large variety of data;
- the data warehouse structure allows it to adapt to changes.

The process of building and using data warehouses is known as data warehousing, and it involves integrating, filtering and consolidating data.

Data warehouses contain different types of data: detailed data, summary data, metadata. The detailed data contains data that refers to recent events, and due to the large amount they require strong computers to manage and process them. The data warehouse should only contain data that is useful and necessary for different analysis areas. The summary data are used more often and are already an analysis and synthesis result of information required by decision systems. Metadata, as a term, can be translated by "data about data" [5], representing a solution for grouping information regarding the data warehouse and the associated processes. There should be no confusion between metadata from the operational environments and the metadata from the data warehouses. While the metadata from the operational systems are almost as important as documentation, the metadata from the data warehouse have a more important role as documentation. The metadata from the two environments are used by different types of subjects, as shown in Figure 1. The metadata from the operational systems serve the IT experts, while the metadata from the data warehouses serve the analysts involved in decision support systems, as they require complete information regarding the way of using the data warehouse, information provided by the metadata.

Another entity similar to the data warehouse is the data mart, which has lead to quarrels between scientists, whether they mean the same thing as data warehouses or not. The data mart is not equivalent to the data warehouse, it is a collection of data by areas of interest, according to the needs of a certain department of the organization. There is a data mart for the financial part, a data mart for the marketing part etc., these data marts being almost totally independent on each other.

Data types are of two different kinds: dependent and independent. A dependent data mart is the one that uses the data warehouse as a source, and the independent data mart is the one that uses its own applications as sources.

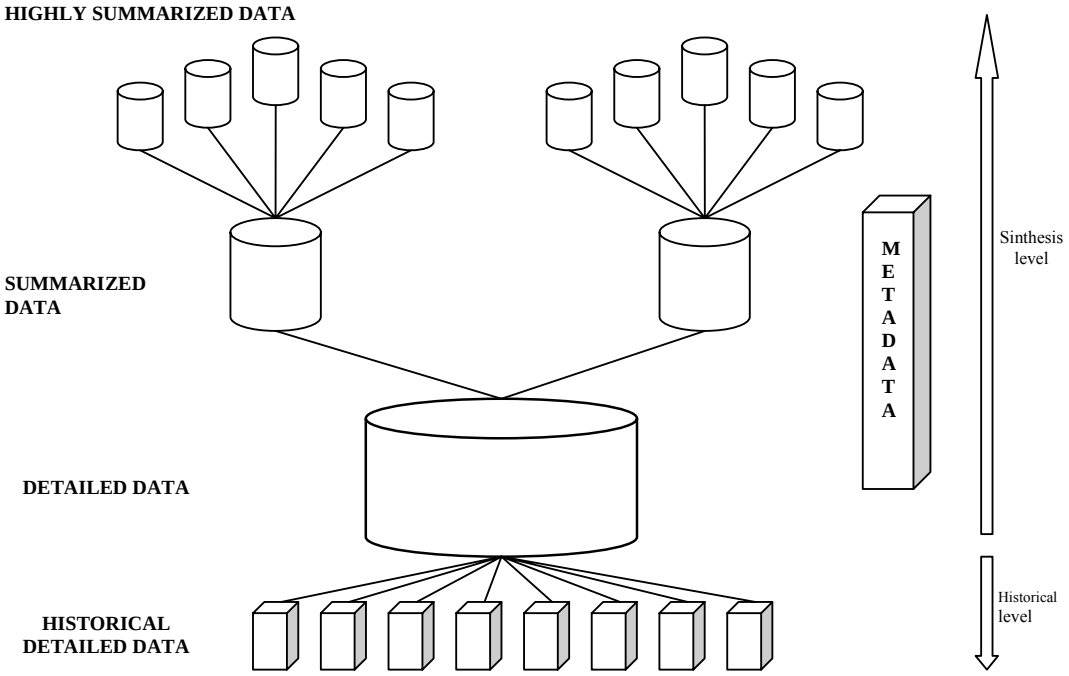

Figure 2. The detail leveles of data within data warehouse (Source: [7])

Dependent data marts are formed after loading data from the operational system in the organization's data warehouse which will be divided in smaller units named data marts, and their dependency is determined by the very fact that they are derived from the data warehouse.

Independent data marts are less stable than the dependent ones, and because of their deficiencies they are not taking action until there are several independent data marts within the organization. As organizations are developing in time, at some point one has to deal with many data marts, each of them requiring data from the operational data base, this fact proving to be quite expensive.

There are several detail levels within a data warehouse (Figure 2) [8]. The inferior level is the level of historical detailed data, and then there is a level of detailed data, followed by a level of summarized data, which refer to data marts, and finally, a last level, of highly summarized data. Data from the data warehouse is provided by operational systems. Important changes of data appear when they pass from the

operational level to the data warehouse, and as data get older, they pass from the detailed data level to the historical detailed data level. As data is being summarized, they pass to the summarized data level and afterwards to the highly summarized data level.

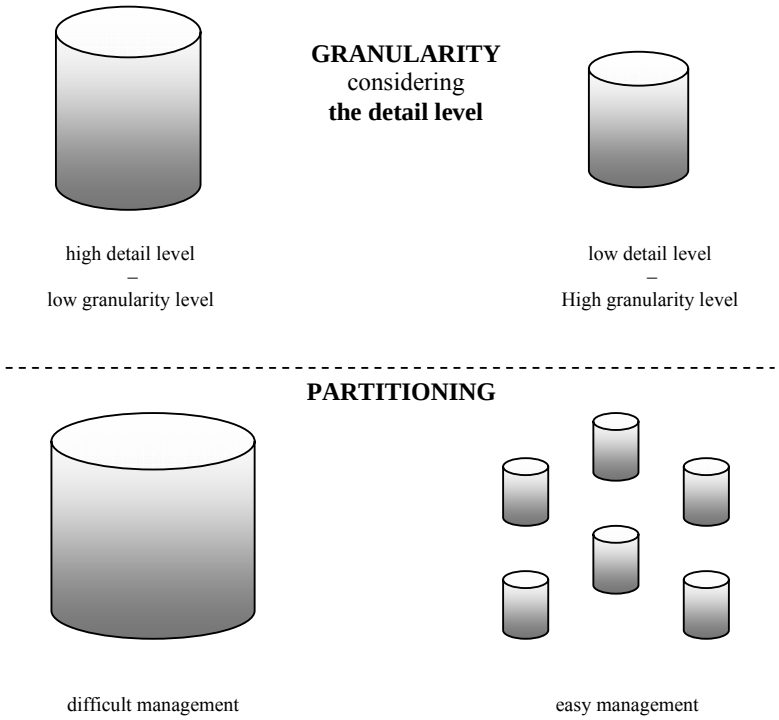

Figure 3. Essential concepts of data warehouse: granularitaty and partinioning (Source: [8])

Granularity is an essential concept regarding the designing of data warehouse, referring to the level of detail or summarizing of data units from the data warehouse (Figure 3), and the granularity level is disproportional with the detail level. The high level of granularity of data from data warehouses is the reason that the designer requires time and resource for detailing the data.

The partitioning of detailed data involves the division of data in small physical units, which can be easily managed. The reason for data partitioning is that some tasks, as restructuring, indexing, reorganizing, are difficult to fulfill when data is organized in larger physical units. Dividing data can be made on different criteria, such as data, area, location, department, depending on the designer, but a criterion should almost every time be the calendar date [8].

Data warehouses can be very useful to various categories of deciders, and the most important ways to benefit from the data within the warehouses are online analytical processing (OLAP) and Data Mining techniques. The OLAP technology refers to the possibility of aggregation of data in a warehouse, being able to filter the large amount of data to obtain useful information for the decisional process within an organization. According to specialists, an alternative term for describing the OLAP concept would be FASMI (Fast Analysis of Shared Multidimensional Information). The

essence of each OLAP is the OLAP cube, also known as the multidimensional cube composed from numeric facts called measurements, categorized by dimensions [4]. These measurements are obtained from records in the relational databases tables. The outcomes of user requirements can be achieved by dynamically traversing the dimensions of the data cube, on a high or detailed level.

OLAP systems have the following properties [5]:

- multidimensional data view;
- intensive evaluation capabilities;
- timeline orientation (time intelligence).

The multidimensional perspective upon data refers to the capacity of integrating several aspects of the company's activity from different points of view: time, location, products, money, persons etc. Each dimension may have several levels: the temporal dimension can be divided in years, month, trimesters, seasons etc., the geographical dimension can be divided in hemispheres, continents, countries, areas, cities etc. The concept of *dimension* is being used and understood as *aspect*, the dimension being completely independent and being measured with all the values of that certain dimension. The measuring units are possible data summarizing criterion and the levels of a dimension form a hierarchy that can also provide data summarizing criterion. The multidimensional perspective is called data hyper-cube, by extension of the tri-dimensional cube to the n-dimensional cube or hyper-cube.

Data Mining technologies, due to their characteristics, are very suitable to analyze large amount of data. Data Mining is aiming to discover patterns within data sets, while other analytical technologies, such as queries, statistical analysis systems are not able to, and OLAP tools are based upon verifications, which prove to be limited.

The collecting of data that reflects an organization's activity has become vital in order to achieve competitonal advantage. The medium and large companies have made investments into computer based systems that collect data and are able to manage very large data bases. The main task these systems have to successfully fulfill is knowledge discovery that follows after the reasoning upon the information that results from the collected data.

Data Mining technologies can accomplish the following tasks:

- prediction – future values of variables we care about can be acquired by finding patterns in examples and developing a model;
- classification – finding a function that classifies the records into discreet classes;
- relation detection – allows searching for the most influent independent variables;
- explicit modeling – describing different variable dependencies through explicit formulas;
- clustering – allows identifying similar record groups that are different from other records not in the group. It is often needed to also identify the variables that lead to obtaining the best clusters.

Patterns established by Data Mining technologies that are generated using prediction techniques prove themselves to be highly important throughout the decision-making process because they bring to light various aspects that can lead to an improvement of the decision-making process, from an efficiency point of view as well as from a time-consuming one.

The large amount of data is outrunning the human processing capacity, also in order the decisions to be correctly grounded, systems that are using Machine Learning technologies are required. These systems allow the discovery of patterns at the very level of unprocessed data, providing different results that can be used within the decision support systems but can also be used by the human analyst.

The Data Mining process consists in four important phases: data collecting, data preparation, pattern discovery and pattern analysis. Data collecting phase involves overtaking data from different sources, and considering that this data could be heterogeneous, the preparation phase will normalize the data and represent it in structures, in order to facilitate the data use. The data identified after certain characteristics during the previous phase is extracted and afterwards formatted, so the data will be represented in the form that the Data Mining application requires. The discovery of new patterns follows from applying Data Mining technologies upon the selected data.

Nowadays, computer based information systems that are using Data Mining technologies are able "to learn" from the previous behavior of the considered elements, and based on the knowledge following from the "learning process" they are able to make hypothesis which they will be testing. The knowledge that proves to be valid and useful can be integrated within the decision support systems, in order to be useful to the decision makers and assist them in making the right decision.

BIBLIOGRAPHY

1. Andone I., Mockler R., Dologite D., Ţugui Al., *Dezvoltarea sistemelor inteligente în economie*, Editura Economică, Bucureşti, 2001
2. Cojocariu, A., Stanciu, Cristina-Ofelia, *Informatică de gestiune*, Editura Eurostampa, Timişoara, 2008
3. English, L.P., *Improving Data Warehouse and Business Information Quality,* Wiley Computer Publishing, 1999
4. Ganguly, A. R., Gupta A., *Data Mining Technologies and decision Support Systems for Business and Scientific Applications*, Encyclopedia of Data Warehousing and Mining, Blackwell Publishing, 2005
5. Graz, P., Watson, H., *Decision Support in the Data Warehouse*, Prentice Hall, Upper Saddle River Publishing, 1998
6. Hamilton, H., Gurak, E., Findlater, L., Olive, W., *Knowledge Discovery in Databases***,** University of Regina, Canada, 2002, http://www2.cs.uregina.ca/~hamilton/courses/831/notes/ml/dtrees/c4.5/tutorial.html
7. Inmon, W.H., *Building the Data Warehouse, 3rd Edition,* Wiley Computer Publishing, USA, 2002
8. Inmon, W.H., *Using the Data Warehouse,* Wiley Computer Publishing, USA, 1994
9. Lungu, I, colectiv, *Sisteme informatice – Analiză, proiectare şi implementare*, Editura Economică, 2003, Bucureşti