# DATA WAREHOUSE AND DATA MINING – NECCESSITY OR USELESS INVESTMENT

Associate Professor, Ph.D. Emil BURTESCU

University of Pitesti, Faculty of Economic Sciences
emil.burtescu@yahoo.com

***Abstract**: The organization has optimized databases which are used in current operations and also used as a part of decision support. What is the next step? Data Warehouses and Data Mining are indispensable and inseparable parts for modern organization. Organizations will create data warehouses in order for them to be used by business executives to take important decisions. And as data volume is very large, and a simple filtration of data is not enough in taking decisions, Data Mining techniques will be called on. What must an organization do to implement a Data Warehouse and a Data Mining? Is this investment profitable (especially in the conditions of economic crisis)? In the followings we will try to answer these questions.*

***Keywords**: database, data warehouse, data mining, decision, implementing, investment.*

***JEl Classification: M15, D80***

In the last decade we assist at an explosive growth of our capacity of generating and collecting data. The progress made in collecting data, spreading the use of bar codes for most of the commercial products and computation of business and government transactions have flooded us with information. It is being discussed more and more about the ocean of data, which in specialty literature is called "ubiquitous". This term doesn't have an exact translation or a very accurate meaning. It refers to data that follows almost invisible the day after day life of modern man. The origins of the data are different, and their existence is almost unnoticeable for human eye. They come from the most diverse sources, from the most simple devices such as electronic washing machines, microwave ovens, digital cell phones to complex databases concerned with population, health etc.

## 1. DATA WAREHOUSE

The data warehouse concept has its origins in the early 60's when following the collaboration between a group of companies and a university in the United States were introduced new terms such as *dimensions* and *facts*.

The role of Data Warehouse was obviously marked by the year 2000 once with the advent of applications accessible to all consumers.

Data Warehouse represents in fact a response to the developers of IT society dynamics. There are two premises that led to the emergence of data warehousing:

1. Economic premises.
2. Technological advances.

The premises are in close relationship with economic and market dynamics, namely: **globalization of trade**; **the dramatic sharpening of competition**; **spectacular shortening of products' cycle of life due to technologic dynamics and imposing of extremely high quality requirements**.

The economic premises are: **the increase of computing power** and **low prices**.

Given the above, we can draw the following conclusions: informatics systems exist; data can be accessed from anywhere; the need for information is acute; great computing power; large and cheap storage capacity and available software tools. So there we have all the premises for implementing a Data Warehouse. Data Warehouse is a collection of designed data for the fundamentals of management decision. Data Warehouse contains a great variety of data that present a coherent image of business conditions of a company at one point in time.

A Data Warehouse like system is used to monitor the activity of organizations by generating reports in standard formats for analyzing issues related to the work in the organization and based on this analysis taking coordination decisions is made.

In a Data Warehouse Data Mining operations are particularly made. One can say that Data Mining instruments use raw materials supplied by the Data Warehouse.

Basic features of a Data Warehouse are:

1. It is focused on daily operations and transactions. It focuses on the data used in the analysis on which decisions are taken.

2. The basic operation is to add data. The data is not deleted and not at all overwritten. A log of data is maintained.

3. The system is integrated. Data is collected from different places (operating systems, databases, files etc.), is transformed by bringing the data to the representation format from the Data Warehouse and centralized in the system.

4. The integration of data represents the most important issue in the construction of a Data Warehouse.

The necessary costs for creating and maintaining a Data Warehouse are divided equally into the following categories:

- Required hardware systems and data storage systems.
- Software needed for extraction, processing, storing and analyzing data.
- Professional services.

Building a Data Warehouse is a complex work and it is addressed particularly to experienced professionals.

A basic Data Warehouse is made out of the following levels with their own structure and functionality (figure 1).

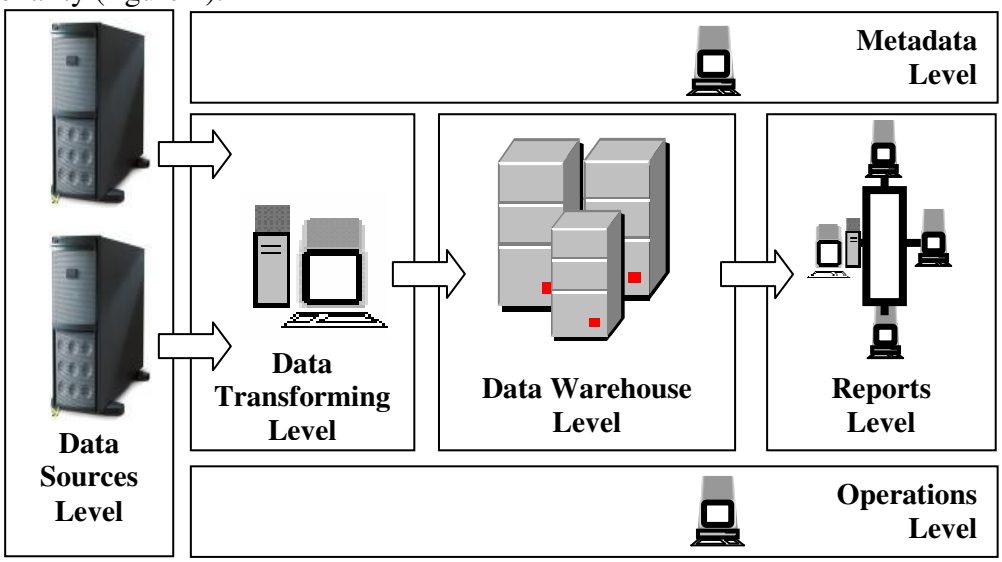


**Figure 1. Data Warehouse structure**

**1. The level of data sources**

The level of data sources refers to any information, usually electronic, that must be collected and stored in the Data Warehouse. It uses as data sources: mainframe databases (DB2, IBMS,

VSAM, ISAM, Ababas etc); client-server databases (Oracle, Informix); small databases, from PC's (MS Access); spreadsheets and other types of files.

**2. The level of data processing**

This level deals with bringing the collected data to a standard form. The transformation consists in bringing them to the internal format of the Data Warehouse. Special applications are used such as ETL (Extract, Transform and Load).

Data transformation may require operations of "cleaning" data (accuracy, consistency, totals, subtotals etc).

**3. Data Warehouse level**

Data Warehouse level is actually dealing with data storage. For storage relational databases are used. Data is retained for determined periods of time.

**4. Reporting level**

The level of reporting is the one that does the analysis and generates reports for monitoring the organization's activity. The generating of reports can be done using specific tools: Business intelligence tools; Executive information systems; Enterprise Information Systems; OLAP; Data Mining and KDD (Knowledge Discovery in Databases), that uses statistic analysis techniques, "form" recognition to discover correlations, rules, knowledge etc.

**5. Metadata level**

Metadata includes administration information of the Data Warehouse (date of last update, number of connected users etc.)

**6. Operations level**

This level has as main purpose loading data in the Data Warehouse, but also manipulating and extraction of data. The second purpose is represented by the user management, security, capacity but other administration functions as well.

Three types for the implementation of a Data Warehouse are known: **analytical**, **for standardization of reports** and **for homogenization and consolidation of data**.

**Analytical**. This is very complex and requires a lot of attention from design to implementation. It is mainly addressed to the analysts who can interpret the data.

**For standardization of reports**. These are the most common types known on the market (80%) because they are easy to interpret. Because of the fact that they are based on developing standard reports, these have appeared long before Data Warehouse.

**For homogenization and consolidation of data**. These are the most complex and combine multiple sources of information in order to fit the data.

Typically Data Warehouses double their size the first 12 up to 18 months.

## 2. DATA MINING

Another very powerful tool, along with the Data Warehouse, that is available to assist organizations as support to taking decisions is represented by Data Mining techniques.

Data Mining is a "deeper search" in the source data. The source data means both the data from the Data Warehouse but also other data categories. Data Mining, also known as "knowledge discovery in large databases" is a modern and powerful IT instrument that can be used to extract useful information but still unknown.

Data Mining in many cases involves data analysis in large data deposits - Data Warehouse. Data Mining is the process of extraction knowledge from databases (Data Warehouse), knowledge that was previously unknown, valid and in the same time operational.

Unknown data = new, unexpected, surprising;

Valid data = extracted knowledge must be translated and applied in practice;

Operational = operational knowledge to be obtained in a short time.

By unknown knowledge we understand extraction of new knowledge that are unanticipated and sometimes even surprising. Extracted knowledge must be translated and applied in reality. Also they must be made operational and obtained in a short time.

Data Mining differs from other data processing for data analysis, such as data query, reports, OLAP etc.

Information obtained through Data Mining techniques can be predictive or descriptive. Predictive information is used to describe an event, such as the possibility of fraud.

Using Data Warehouses and data extraction from warehouses using Data Mining techniques will give the organizations that use them a clear advantage over the competition.

Implementing Data Mining techniques, also as in the case of Data Warehouse, must be done by specialists in order for the results to be the ones expected.

Operating principle in the processing of data mining is illustrated in next figure (figure 2).

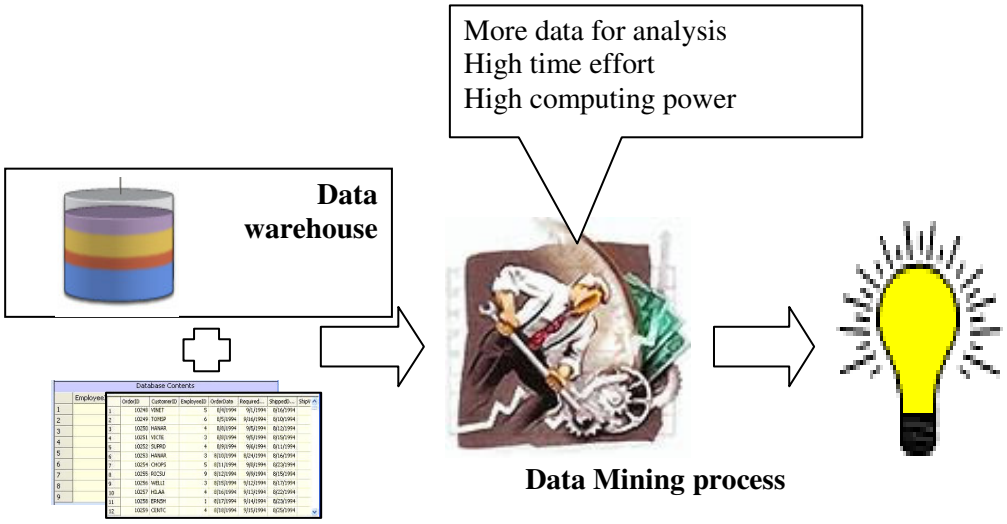


**Figure 2. Data Mining process principle**

In detail, a Data Mining process consists of the following steps (figure 3).

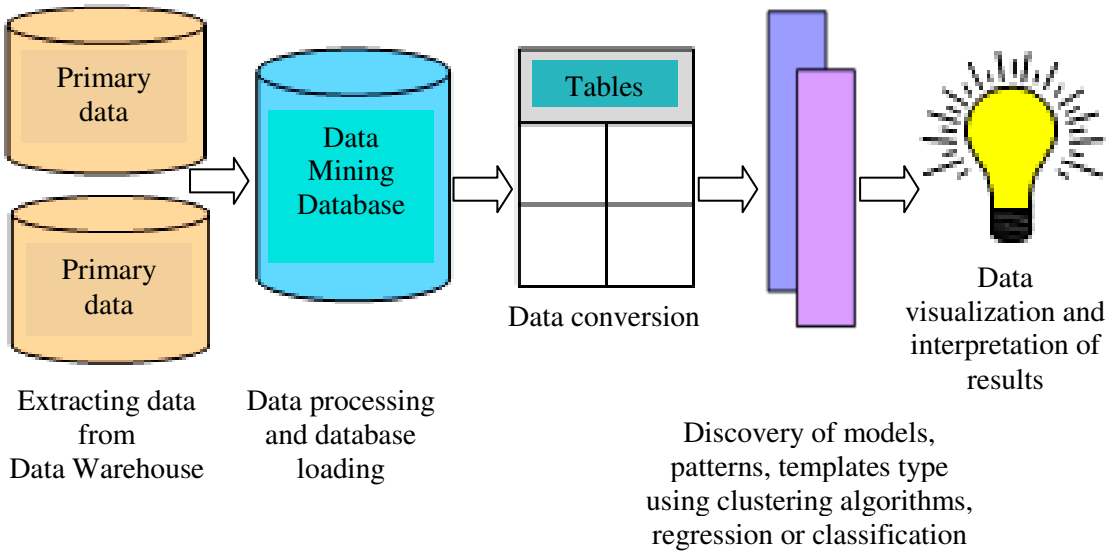


**Figure 3. Data Mining process detail**

The areas where Data Mining is used are multiple. They are used especially in the following directions:

- sales analysis and establishing customer behavior;
- research (medical research);
- fraud, electronic fraud and (cyber) terrorism ;
- risk analysis.

Data Mining tools use technologies of Artificial Intelligence (AI) to process and extract data. Lately, Artificial Intelligence solutions are more and more present in the offers of the companies who offer data analysis software for business activities, thereby we have the Business Intelligence (BI) concept. Using a Business Intelligence solution is suitable both for the management departments and for the other departments. Each department can use specific capabilities of a Business Intelligence solution. Architecture of Business Intelligence solutions is illustrated in next figure (figure 4).

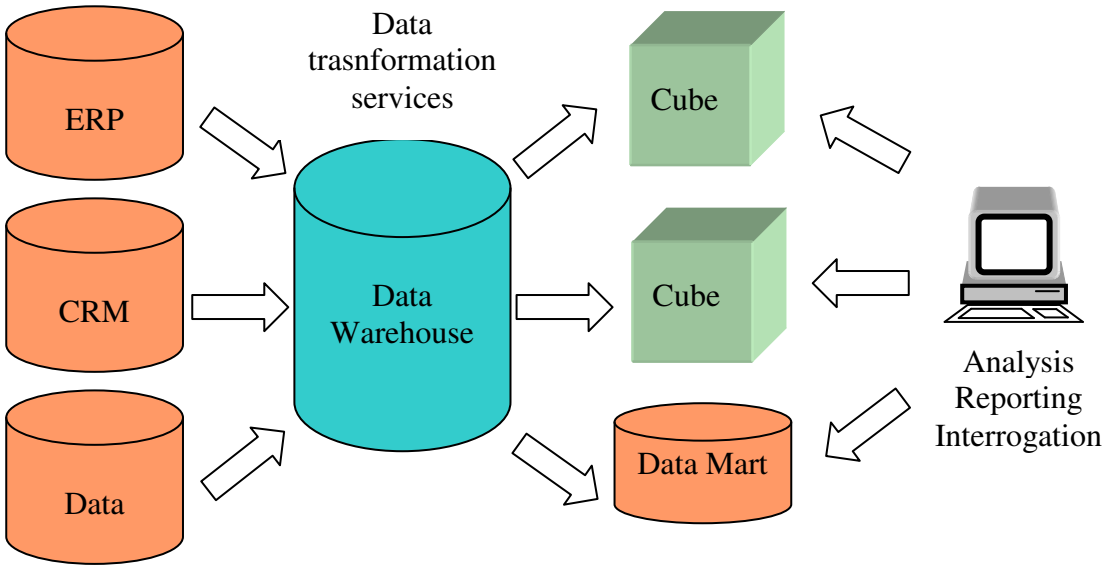


**Figure 4. Architecture of Business Intelligence solutions**

The benefits are:

- Financial departments will use Business Intelligence to analyze data and generate reports, financial statistics and to take important financial decisions based on the information.
- Operations department can anticipate requests, estimate stocks and can make the production process more efficient.
- Marketing and sales department can analyze the response to promotions and marketing campaigns, can estimate sales and the analysis of market behaviour.
- Freeing IT departments from tasks.

Components of Business Intelligence solution are: Windows Server 200X, SQL Server 2000 Analysis Services, Microsoft SQL Server Accelerator for Business Intelligence, Project Server 2000, SharePoint Portal Server, Data Analyzer, Map Point and connection customers.

Commercial Data Mining tools can be found at:
http://www.oracle.com/technology/products/bi/odm/odminter.html
http://www.sas.com/
http://spss.com/clementine
http://www-306.ibm.com/software/data/iminer/

**Implementing the system**

To design and implement a Data Warehouse and Data Mining it is necessary to go through some stages. We will discuss about Data Warehouse, understanding the existence of Data Mining.

There is no strict methodology to design and implement a Data Warehouse. In the following we will define the important stages without concentrating ourselves on technical elements.

**Stage 1. Implementation decision**

The first question one has to answer when someone wants to implement a Data Warehouse is: "Do I really need this?". The answer can be multiple some of them being clear and others ambiguous. Ten years ago probably, especially in our country the answer would have been: "No". A great part of the decision would have been taken due to the ignorance of the used terms-Data Warehouse and Data Mining. Nowadays the answer must be emphatically "Yes". A decisive argument in the foundation of the response is given by the sharpening of the competition on the market. Companies are forced to explore a huge volume of data in a very short time in order to take a decision. A second argument is given by the companies that study the IT market. Gartner Groups Inc. was making in 2009 the following predictions for the upcoming years:

- by 2012, due to the lack of information, processes and tools, more than 35% from the total of 5000 companies worldwide will have problems in taking optimal decisions for the company's activities.
- by the year 2012 the companies will allocate 40% of the budget for Business Intelligence.
- in the late 2010, 20% of the organizations will rule analytical software applications as a Business Intelligence component.
- by 2012, one third of the analytical applications will be delivered through Web applications (web application hybrid).

**Stage 2. The analysis of the existing economic system**

At this stage a strategic analysis referring to the company's assets and the ways of achieving these objectives, and also an analysis of the existing informational system must be done. From the existing economic system we will retrieve data which consequently we will use to help management factors in taking a quick and correct decision. If the collected data is not correct and the decision will be incorrect - *garbage in garbage out*. In many cases, from personal findings the managers expect certain errors to be corrected by the IT personnel.

**Stage 3. Analysis of the existing IT system**

At this stage 2 things must be established:

- if the existing IT infrastructure can be used (at least partially to reduce costs) to implement Data Warehouse and Data Mining;
- whether the organization has a well defined and formalized security policy (in conjunction with step 3).

Hardware components that will form a new architecture cost one third of the total cost so I have to see if I can lower this cost. Ideal would be to find as many solutions to reduce initial costs, by relocating the existing computer systems, following that these to be replaced by the ones initially indicated. We should not exaggerate, an equipment with minor problems will create big problems in such a system.

In a Data Warehouse the important data about the organization is stored in one place. This involves increased risks for the organization. Building a Data Warehouse must be done very carefully because having all the data of a company focused in one single point data security problems might appear. The steps that follow are much more complex gaining technical meanings.

**Stage 4. Data Warehouse architecture design**

Taking as starting point the existing architecture we can define the future architecture. The new architecture must be designed in such way that it will have as little impact as possible on the existing system but a further development to be allowed [3]. We define physical and logical configurations, the data, the necessary applications and financial and architectural support are designed. Basically the main components must be designed: the data stores, ETL (Extract, Transform and Load) system and the front-end applications.

**Stage 5. Selection of the technological solution**

At this stage we will identify the implementation tools for data and applications, and the tools needed for technical and architectural support. Selecting these tools must take into account the structure and the complexity of the Data Warehouse. According to their function, these tools can be classified into the following categories: Transforming and extracting time, data cleaning, data loading and refreshing, data access, security providing, version control and configuration management, database management, data backup and recovery, disaster recovery, performance monitoring, data modeling, metadata management [3].

**Stage 6. Development**

At this stage the elements designed in the data stores are designed, ETL system (including data quality system and metadata) and the front-end applications.
We define the detail level of design for every operation that has as purpose data extraction.

**Stage 7. Testing and implementation**

At this stage, all designed and built elements are put together forming a system, and are tested and implemented thereafter.

**Stage 8. Operation and maintenance**

This is actually the last stage, with equal duration of the Data Warehouse life. It involves using Data Warehouse for the purposes for which it was designed and also for periodical application on Data Warehouse of some test and maintenance operations, evaluation and security assessment. All these stages are done one at a time, when one is finished the other starts-waterfall methodology.

## 3. WAYS OF IMPLEMENTATION

To be able to define a strategic plan for initial implementation and further development it is necessary to know the financial resources, time and specialists of the company. In the case of financial resources an implementation at a lower level for the Data Warehouse and Data Mining must be done. Depending on the time available one can choose for the application development his own personnel or can choose the COTS products (Commercial On The Shelf). This option must be put together with the level of the company's specialists. If the training of the specialists is high, then the implementation can be done, but otherwise this operation must be outsourced, with the problems that can appear starting from this point.
Creating and developing a Data Warehouse can be done starting from the following three stages:

1. Implementation from scratch.
2. Improving the existing system.
3. Converting an older system.

Implementation from scratch has the advantage that I do not need an audit of what we have in the organization but it implies higher costs. For operating safely it is advised to start from scratch. Instead upgrading the existing databases or hardware and software will reduce costs, but will also create "fear" in later use. If something goes wrong later, then I would probably ask myself: "Why haven't I changed the … too?". With a limited budget I will probably be forced to use open-source applications.

With a budget that is over the calculated average we will have a new and reliable system, with increased computing power. It would be best to have all the amount of money in order to make all the assembly powerful.

An option to reduce costs would be outsourcing certain services. Outsourcing any service must be done by weighing three factors: money, performance and security. With less money I can get more performance but I need to share data with an outsourced organization.

## 4. RETURN ON INVESTMENT

For this we have to answer the question: "How do we measure success?". Displaying new results will be a gained point in understanding the necessity of building and exploiting the Data Warehouse and Data Mining.

Time factor will also be essential. Making a report in a very short time will put another brick in the building of the system.

But most important is the fact that all the system can be used in assisting decisions, this time disposing the results of a large data volume, so with a higher precision.

Finally, an analysis with the specialists in the company or outside the company will highlight the benefits of using Data Warehouse and Data Mining.

The most important thing is represented by the acknowledgement of resources owner over the necessity of this type of investment.

There are Data Warehouses consisting of databases containing between 1 and several tens of terabytes. Creating such a warehouse costs around 3 million dollars. What organization would invest so much money without obtaining profit?

Now we have more clearly - it is time to improve the organization databases by creating the Data Warehouse and using Data Mining tools on these and on other data in the system.

## 5. CONCLUSIONS

For the medium and large organizations with a very good computer system, implementing a Data Warehouse and of course a Data Mining are absolutely necessary.

Even if it seems an easy task at first sight, implementing a Data Warehouse proves to be a challenge for the specialists. Lately, more and more commercial applications come with Data Warehouse facilities implemented. Very often, the organization implements a Data Warehouse and thinks the issue is solved. But things are not as easy as they seem. Probably many companies have specialists who can read a standard report given by a Data Mining tools on a Data Warehouse. But how many companies have analysts who can read and interpret an analytical report or a report for homogenizing and consolidating data? This is where things get complicated. The decision personnel must be assisted by persons who are capable to interpret the data, which consequently will be used in the decision process.

Finally, after the system will be used in assisting the first decisions, the benefits of using a Data Warehouse and Data Mining system will be seen.

## REFERENCES

[1] F. Gorunescu, *Data Mining – concepte, modele □i tehnici,* Ed. Albastră, 2007.
[2] V. Rainardi, *Building a Data Warehouse: With Examples in SQL Server*, Apress, 2008.
[3] http://revistaie.ase.ro/content/42/velicanu.pdf
[4] http://www.gartner.com/it/page.jsp?id=856714