Building a Data Warehouse step by step

Manole VELICANU, Academy of Economic Studies, Bucharest
Gheorghe MATEI, Romanian Commercial Bank

Data warehouses have been developed to answer the increasing demands of quality information required by the top managers and economic analysts of organizations. Their importance in now a day business area is unanimous recognized, being the foundation for developing business intelligence systems. Data warehouses offer support for decision-making process, allowing complex analyses which cannot be properly achieved from operational systems.

This paper presents the ways in which a data warehouse may be developed and the stages of building it.

Keywords*: data warehouse, data mart, data integration, database management system, OLAP, data mining.*

I Data warehouse presentation

To cope with the challenges of a more and more global market, to withstand a stronger competition, the organizations are obliged to store more and more data, concerning both their own business and competitive environment in which they are operating. This data represents an important resource of the organization and it has an essential role in the decision making process concerning the ways to follow. The company performances are as good as data is more relevant and better organized. Data analyzing leads to a better understanding of customers' needs and behaviour, of products and services offered to them, and even of organization itself. Data in the operational systems of the company, which refers to specific activities and covers short period of time, cannot offers the proper answer for making stategic decisions. This thing leads to the building of analytical systems, based on data warehouses, in which information are integrated from different sources, both internal and external. Data is organized depending on the major subjects of the organization and is retained on long periods of time (5 – 10 years).

Data warehouses and analytical systems based on them bring out the informational potential of date stored during years. On the basis of integrated, systematised and consolidated date it is possible to take strategic decisions based on complex analyses concerning the company activity and performances, data correlations and forecasting future trends.

Data warehouses are the result of the interference between economic environment and advanced information technologies. The economic environment is more and more global, complex, and competitive and requires proper information for making strategic decision.

II. Data warehouse building

Data warehouse development is a continuous process, evolving at the same time with the organization. Untaking into consideration this aspect may lead to loose necessary information for future strategic decisions and competitive advantage.

A data warehouse implementation represents a complex activity including two major stages. In the first stage, of system configuration, the data warehouse conceptual model is established, in accordance with the users' demands (data warehouse design). Then data sources are established, as well as the way of extracting and loading data (data acquisition). Finally, the storage technology is chosen and it's decided on the way the data will be accessed.

II.1. Data warehouse development

The data warehouse development must take into consideration the users' requirements concerning reporting and analysing. Otherwise, the data warehouse will become a "data jail" which will hide important information

that users need [HYPE00]. To do that: one or more business processes are selected for modeling, granularity is established for each business process (the atomic data level), dimensions that will be applied to every row in the fact table are established, measures that will populate the fact table are selected.

To reach these goals, a detailed analysis has to be made for all the databases containing data that can be extracted with the view to populate the data warehouse. It's very important to correctly understand the data from different systems within the organization and the relations between them. The management of these relations during data warehouse loading is essential.

Another important step is the identification of equivalent entities in different operational systems. The same information may be stored in fields having different names. Also, fields having the same name may contain different data.

There are two **approaches** a data warehouse may be developed: top-down and bottom-up.

The top-down approach is a mature one and it is used when the technology and the economic problems are well known. This approach achieves the sinergy between the business subjects and provides a single version of the truth. It is a systemic method which minimizes the integration problems, but is expensive, of long standing and has a low flexibility.

The top-down approach suits the vision of Bill Inmon, who considers that the data warehouse must respond to the requirements of all the users in the organization, and not of only a certain group.

The bottom-up approach is a fast one and is based on experiments and prototypes. It is a flexible method that allows the organization to go further with lower costs, to build independent data marts and to evaluate the advantages of the new system as they go along. However, there could be problems when trying to integrate the data marts in a consistent enterprise data warehouse. Since in the first iteration data definition is not made by consulting more business lines, the solution could be rejected by the next business line involved.

This approach suits the vision of Ralph Kimball, who considers that the data warehouse have to be easy understood by the users and provide correct answers as soon as possible. This approach starts from business requirements, while the top-down approach has in view data integration and consistency at level of entire enterprise.

The two approaches may be combined to benefit by the advantages provided by each of them. From software engineering point of view, one of the following **methods** can be used:

- *waterfall approach*, which requires a structured and systematic analysis at each step, before going forward;
- *spiral (iterative) approach*, which allows fast generation of more and more developed functional systems.

The most adequate method for developing a data warehouse is the iterative one. In this approach, more iteration is made, a new version resulting at every iteration. The business subjects are approached one after the other. The method provides a scalable architecture and answers the informational demands of the whole organization. It also allows an efficient management of the users' requirements and reduces the possible risks.

Unlike other software systems, development of a data warehouse in a single large project (the "big bang" theory) has low chances to succeed. As Bill Inmon said, the organizations need an evolution, not a revolution [INMO05]. Through this iterative approach, tangible results are obtained at short periods, with low costs.

II.2. Stages of building data warehouses

The stages of building a data warehouse are not too much different of those of a database project. Specific to data warehouses is the fact that they are built through an iterative process, which consists in identification of business requirements, development of a solution in accordance with these requirements and implementation of data warehouse architecture.

The following *stages* are to be completed to develop a data warehouse: development of a feasibility study, business line analysis, data warehouse architecture design, selection of the technological solution, planning the project iterations, detail designing, data warehouse testing and implementation, deployment and roll-out.

1. Development of a feasibility study. This stage starts with a strategic analysis, including the evaluation of organization business lines. More business lines will be implemented in the long term, but in the short term it's better to choose the subject having the greatest strategic priority. The feasibility study defines the activities, costs, benefits and critical factors for the future system success. The data warehouse will be built in a number of designing, developing and refining iterations according to the tactical and strategic business requirements. In this stage, both short term and long term strategies are pursued, the immediate and further costs are identified, so that a proper budget plan could be fulfilled.

The feasibility study is presented to some important managers within the organization to determine them to become project sponsors. The greater the sponsors' authority is, the greater the chances to succeed are. It's a good idea to co-opt a sponsor from the business area and a sponsor from the IT department.

The roles and responsibilities for all the people involved in the project must also be established in this stage. This thing leads to a clear establishment of the relations between the team members, to a better understanding of the project and to an improved communication between the participants.

2. Business lines analysis is an important stage in the data warehouse development cycle. Its main purpose is business understanding and business requirement identification. Usually, the development of an operational system has only one sponsor and the system will have a well outlined group of users with a clear vision on its requirements and functionalities. On the contrary, the users of a data warehouse form a heterogeneous group and they will formulate varied requirements, which only partly can be foreseen in data warehouse development stage.

The following *goals* must be achieved in this stage: achieving a global view on organization's activity and users' requirements, establishing the data warehouse scope, identifying the business directions and purposes, establishing the priorities of users' requirements, establishing the necessary data for solving the requirements, defining the process for data warehouse iterative population and for data validation according to the business requirements.

All the necessary information for identifying and understanding the business requirements are obtained from debates and interviews with the final users. The subject of these debates consists in identifying the information that can help the users, and not what they think that have to be stored in the data warehouse. These discussions have to be focused on the users' goals and the ways they take the decisions. Therefore, the topics of discussions will be about the following *problems*:

- the mission and functions of the group or department, the key performance indicators used and the critical factors for success;
- the products manufactured, the providers and customers of the organization;
- the information systems and applications used; the manual transformations necessary to get the data that is not available;
- the detail level of the data and the frequency of receiving the requested reports; the way these reports answer the demands, concerning both their content and interval of getting them;
- the period the data must be retained.

At the same time, the team must talk with the administrators of the operational system for identifying the data needed to answer the business requirements. Therefore a preliminary data source audit has to be made. The data is evaluated from the point of view of availability, quality and costs necessary to load it into the data warehouse. To achieve this purpose, some *analyses* regarding the following are to be done:

- the current technological architecture of the organization: computing equipments, operation systems, database management systems, networks, development, tools for communication and data access etc.;
- the relations between the different systems within the organization and their level of integration;
- the available documentation, its accuracy and up-to-dateness;
- data quality and possible extracting tools.

On the basis of information got from these interviews, the business subject priorities are established, depending on their relative importance, the costs and feasibility of the necessary data. This priority list is used to establish the scope of the first and the subsequent iterations of data warehouse development.

3. Data warehouse architecture design. The architecture is the logical and physical foundation the data warehouse is built on. In this stage, the components that already exist in the organization are identified, as well as those that are missing and have to be built or acquired for data warehouse architecture completion.

The data warehouse architecture must be designed so that it can allow further development with a minimal impact on the existing model [MSDN06]. It must present the relational and multidimensional databases used for the data warehouse and the data marts, their localization and interconnection, as well as the access tools.

First, the data warehouse *logical architecture* is defined. This is the configuration of the required data collections: a central repository storing the data of the entire organization, an optional operational data store, one or more data marts and one or more metadata repositories.

Once the logical configuration is defined, there must be designed the data, application, technical and support architectures needed for data warehouse implementation. The data warehouse must be optimized according to the users' demands. To achieve this purpose, a carefully analysis has to be made concerning the requirements of these four architectures.

The data architecture has the purpose to organize the data sources and collections and to define the quality and management standards, both for data and metadata. This general model presents the business areas and their relations. One of them will be selected as a pilot area, the one that data warehouse development will start with.

The application architecture presents the software components that provide the implementation of the business functionality within the data warehouse, as well as the data transfer from its source to users, that is data extracting, cleaning, transforming, loading, refreshing and accessing.

The technical architecture provides the proper infrastructure for data and application architectures. It includes the server, network, hardware and software components for connecting and communication, and the users' workstations. The technical architecture must respond to the requirements of scalability, performance, availability, stability and security. It must be robust, reliable, flexible, extensible and parallel [ORAC02].

The support architecture includes tools for backup/recovery, archiving, performance monitoring, as well as the organizational functions necessary for the technological investment management.

4. Selection of the technological solution. The purpose of this stage is to identify the possible tools for implementing data and application architecture and for providing technical and support architecture functions. There must be selected and acquired the most suitable tools for the business and technical requirements that have been defined by the data warehouse architecture. The instrument selection must take into consideration the technical infrastructure of the organization.

The data volume estimated to be load into the data warehouse, platform scalability and the way it supports the selected software components must be taken into consideration when the hardware platform is selected.

At the same time software components are selected: operation systems, databases management systems, development and analysis tools etc. The data warehouse structure and

size are influencing the software requirements. Therefore, if the data warehouse will have data marts too, besides the relational technology, a client-server architecture is necessary to allow data accessing and multidimensional analysis.

According to their function, these tools can be classified into the following *categories*: data extracting and transforming, 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.

5. Planning the project iterations. The data warehouse is implemented one subject area at a time, depending on the priorities established in the stage of business requirement identification. In this stage, the identified business and technical requirements are refined for leading to the proper detail level for data warehouse development and implementation.

The preliminary analysis made in the previous stages is detailed and the data warehouse scope is reviewed. All the constraints imposed by the source systems are identified and documented. These constraints influence the way the data warehouse will be developed.

The documentation has to be reviewed and approved by the project sponsors, to make sure the management expectations and the established goals will be achieved.

The source system audit must be done too in this stage. If the preliminary audit made in business requirement identification stage had the purpose to offer a complete list with data sources, a general view against the potential data sources must be achieved in this stage too.

Most of the data that will populate the data warehouse proceeds from internal sources, especially from data collections of the operational systems. But also financial data and reports can be used, especially if the queries and reports that will be requested refer to profitability indicators, as well as external data.

The database administrators, system administrators and other IT specialists which administrate the internal systems of the organization have the main role in source system audit. As they become familiar with data quality problems, with the rules of obtaining the reports, they are the most adequate persons to estimate the opportunity of using every data source and to offer the proper information to data extract, transform and load process.

6. Detail designing (data warehouse modelling). In this stage the data warehouse physical model (database schema) is achieved, metadata is defined and data source list is updated to include all the information necessary for the implementation of that subject. The data warehouse physical model must respond to the users' informational demands. The data warehouse schema may be developed in accordance with the relational model, based on data normalization, or the multidimensional one, based on denormalization.

At the same time, the source fields are transformed on the destination fields. There are more possible types of transformation. Thus, a field in the data warehouse can be populated with data from several source systems. This is a natural consequence of the data warehouse integrating role, a feature underlined by every data warehouse definition.

It's possible that a destination field is populated with data proceeding from several source systems, or conversely, one source field may populate several destination fields. The typical example for such transformations is represented by addresses, which can be stored in a single text field or in several fields, one for each component.

It's important to clearly specify all the rules concerning the transformation of the source fields in destination fields, no matter if it's a question of integration, splitting or just movement.

Applications are also built in this stage to offer support to the users in analyzing the data from the data warehouse. There are tools for this purpose to build OLAP cubes and data mining models, but is possible to develop other applications too, such as predefined reports, Web sites or dashboards. The func-

tionality of these specialized applications is directly determined by users' demands.

At the same time, the detail design of the whole necessary procedures is completed and the related documentation is elaborated. The implemented procedures have to provide the following *activities*: data extracting, filtering, transforming and cleaning; data loading and refreshing; system security; data accessing; backup, recovery, archiving; disaster recovery; configuration management; testing; users' training; transition to production; helpdesk; version management.

7. Data warehouse testing and implementation. Once the planning and design stages are completed, the current iteration for data warehouse implementation may start. In this stage, the development and testing environments are established, the hardware and software components are installed and the configuration management process is implemented. Referring to the *specified aspects* of the installed database management system, the data warehouse logical and physical design are finalized:

- the physical design of the fact and dimension tables are finalized;
- the most proper indexes are established, taking into consideration the estimated size of the data warehouse and the queries supposed to be performed;
- a decision about table partitioning is taken, knowing that it's easier to manage a partitioned data warehouse, but its performances are lower.

At the same time, programs are developed or configured for data extracting, cleaning, transforming, loading and periodically refreshing. The initial data is loaded, as well as the technical and business metadata. User interfaces and production reports are developed, sample ad-hoc queries are run against the test database, and the results are validated. The data access is established and the user training is done. Support procedures are implemented for database security, backup and recovery, disaster recovery and data archiving.

After the initial test made by the development team, the final users are involved. They have to use the system as they will do after data warehouse transition to production. This manner makes possible to find out and correct the errors, to identify the requirements for performance enhancement, and to allow the users become familiar with the new system. It's possible that other changes will need to be done after the transition to production, but starting this activity with a proper performance represents one of the project success keys.

8. Deployment and roll-out. In this stage the production database is created and the programs for extracting, cleaning, transforming and loading data are run against the source systems. The users' training is completed, the team for change management is created, and the control procedures for future development cycles are established.

In the operation stage, besides the data warehouse employment by the final users, its maintenance and development are provided too. The IT specialists have to do several specific activities for achieving this purpose:

- *Periodical refreshing of the data warehouse*. From time to time, the new changes made in the operational systems have to be loaded into the data warehouse, so that its users could have the most recent information at their disposal. Usually, this process is performed when the operational systems are not in use and it consists in extracting, cleaning, transforming and loading data into the data warehouse.
- *Computing certain statistical indicators* to pursue the data warehouse evolution, performance and maintenance. Some of these indicators are [VELI05]: number of queries performed in a certain day, average response time, number of users per day, frequency of using data warehouse subjects, average duration of a work session.
- *Evaluation of the database size.* The data warehouse size increases at every loading operation and can cause serious troubles. There are several techniques that can be used for reducing the negative consequences the increase of the data warehouse volume over certain limits:

  - aggregating the detailed data, and then archiving and deleting it from the data warehouse;
  - limiting the storing interval at a certain period of time, and then archiving and deleting the data;
  - deleting the unused data, which can be identified on the basis of statistical indicators regarding the data warehouse.
- *Database disaster recovery.* The information within the data warehouse have a strategic importance for the managers of the organization. That is why a special attention has to be given to disaster recovery procedures.

From this issue it is obviously that the process of data warehouse development is a very complex one, which can be achieved only in an iterative manner, one subject area at a time. A data warehouse development cannot be performed if its goals are not clear and well understood. At the same time, all the users' requirements must be identified and the way they will interact with the data warehouse must be established. Having in mind that, a data warehouse must offer support in the decision making process, only the understanding of all these aspects guarantees the success of the project.

References

[HYPE00] *** *Hyperion Essbase Data Mart Design Approaches*, 2000, www.dev.hyperion.com

[INMO05] W. H. Inmon, *Building the Data Warehouse*, 4th Edition, Wiley Publishing, Inc., Indianapolis, 2005.

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

[MSDN06] *** *Data Warehouse Designs Consideration*, 2006, www.msdn.microsoft.com

[ORAC02] Oracle Corporation, *Data Warehouse Fundamentals*, Student Guide, 2002.

[VELI05] M.Velicanu, *Depozite de date*, Suport de curs, Master "Baze de date – suport pentru afaceri", ASE Bucureşti, 2005.

[VEMA07] M.Velicanu, Gh.Matei, *Database versus data warehouse*, International Conference on Economic Informatics, Bucharest, 2007.