The audit of the Data Warehouse Framework[*]

José A. Rodero, José A.Toval
Departamento de.Informática: Lenguajes y Sistemas Informáticos
Universidad de Murcia (Spain)
jrodero, atoval@dif.um.es

Mario G. Piattini
Departamento de Informática
Universidad de Castilla-La Mancha (Spain)
mpiattin@inf-cr.uclm.es

Abstract

Data warehouses have become the key trend in corporate computing in the late 90s, as they provide managers with the most accurate and relevant information to improve strategic decisions. A specific control system should be established in order to protect this important asset. Having COBIT (Control Objectives for Information and related Technology) as our base, we settle a control system for data warehouse projects. This system consists of a number of control objectives plus a methodology to accomplish the audit process. Auditing data warehouses will check the control system, either assuring that control objectives are met or evaluating the current risks associated with controls' lacks.

1 Introduction

The need to have a separate data base in order to support the decision process was firstly recognized at the beginning of the 70s [Sprague et al., 1996]. Although the term "data warehouse" was not coined by Bill Inmon until the late 80s, by 1985 big telecommunication companies and banks began to build systems whose function was to provide horizontal and global information on the organization to the managers [Watson et al., 1992]. These systems were called EIS (Executive Information Systems) and constitute one of the main precursors of the data warehouses.

The main purpose of the EIS consisted in giving support to managers who intended to learn about the organization, including the study of the main processes and the interactions with the external enterprises. It is demonstrated that its use enhances the decision process, providing an improvement of managerial results [Vandenbosch et al., 1992].

The high risk of this kind of systems was detected in these first projects, as the percentage of failures was higher than in traditional information systems. One of the causes is that these systems are devised to be used by managers, who do not have much spare time for learning complex systems or supporting not justified errors [Preece et al., 1994].

Day by day, the amount of data stored in computerized systems grows spectacularly. However, these vast amounts of data, obtained at a relatively low cost, do not have a reflection on the benefit reported to organizations, especially because they do not provide information [Gardner, 1998]. In fact, an organization may be rich in data but however poor in information. Moreover the growing globalization of the economy is generating a stronger competition in different managerial sectors. In order to survive in this turbulent environment, companies have to be flexible, and respond quickly to all the events happening around them. So, nowadays managers need all necessary information (of course accurate information) just at the right time. They demand new kinds of information systems, giving them not only the automation of the information, but also a true exploitation of it.

This need has a response: the data warehouse. A data warehouse is defined as a collection of subject-oriented data, integrated, non-volatile, that supports the management decision process [Inmon, 1996a].

The importance of data warehouses in the computer market has grown increasingly during the 90's, and today they constitute one of the main trends for the development of information technologies.

Once the data warehouse is operational in a corporation, it becomes a strategic tool for decision making. The first question we have to answer is:

Why is it necessary to fix mechanisms of control and audit?

The reply is based on three main arguments:

[*] Partially granted by CICYT (Science and Technology Joint Committee), Spanish Ministry of Education and Spanish Ministry of Industry, research project TIC97-0593-C05-02.

Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'99)
Heidelberg, Germany, 14. - 15. 6. 1999
(S. Gatziu, M. Jeusfeld, M. Staudt, Y. Vassiliou, eds.)

http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-19/

a) Most of the strategic decisions will be based on the information the data warehouse delivers, and everybody knows the expensive consequences a failure may have on strategic decisions.

b) The development investment is usually very high. According to [Haley et al., 1998], the cost is usually higher than 2 million dollars.

c) Data warehouses are high risk systems, with a contrasted failure rate reaching about 50%.

Therefore a data warehouse is a risky and high valuable asset for the company, and not only for the money that it costs, but for the information that it holds. As far as the management is concerned, one of their main responsibilities is to safeguard all the assets of the enterprise. In order to discharge this responsibility, as well as to achieve its expectations about the data warehouse, management must promote a system of internal control guaranteeing the protection of this asset.

From the data warehouse users' point of view, there must be some guarantees in the process of fulfillment of the minimum quality requirements regarding supplied data. For this purpose, independent accreditation processes are needed.

To comply with both requirements, protection from the management's point of view, and a guarantee for accuracy and certainty from the users' point of view, it is necessary to settle a system of internal control for data warehouses. Figure 1 shows the audit process, which is necessary to verify that the internal control exists and works well. It will check the effectiveness of these controls and their correct and continuous application. The result of the audit process will be an assessment of the risks associated with the lack or misfunctioning of control.

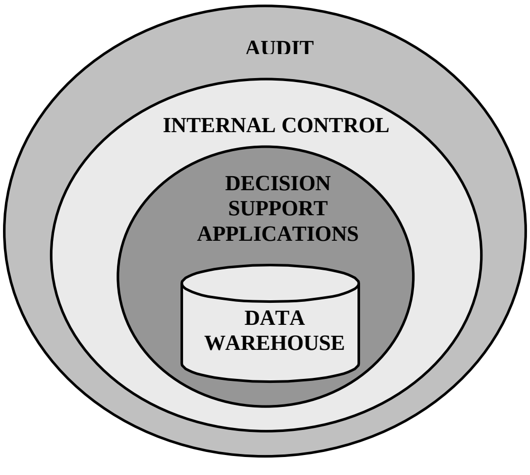

Figure 1: Audit and control of data warehouses

To accomplish the audit of data warehouses, we propose a control system based on COBIT (Control Objectives for Information and related Technology), more specifically on the second version, published in April 1998, by the Information Systems Audit and Control Association (ISACA).

The fact that the data warehouse consists of applications, data, utilities, technology and people, together with COBIT's strong business orientation turns them into an ideal instrument for the purpose we intend. In [ISACF, 1998a], a control objective is defined as "a statement of the desired result or purpose to be achieved by implementing control procedures in a particular IT activity".

Our proposal includes a number of specific control objectives, related to this kind of project, plus a methodology to accomplish the audit process. The main objective is to evaluate the control system, assuring that control objectives are achieved or evaluating the current risks associated with the wrong operation of control.

In [Rodero and Piattini, 1998] we propose the audit of data warehouse from the life cycle point of view, describing a well-accepted life cycle for data warehouses, including the explanation of the main stages and their related control objectives. To complete this task, we approach the audit from the framework point of view in this paper.

Since the definition of both the framework and the control objectives is the most important task in the audit, we will focus our attention on them in next section. A schema for the audit process is presented in Section 3. Conclusions are drawn in Section 4.

2 The Data Warehouse Framework and the related Control Objectives

2.1 Introduction to the Data Warehouse Framework

Data warehouses are not simple systems. Their natural complexity, owing to the kind of problems they are intended to solve, providing business analysts a unified view to information, is added to the lack of a model that defines which techniques should be applied and which is the framework for the development of this kind of systems [Kelly, 1997].

For a given project, the framework must be stable with regard to its philosophy; however, it must be dynamic and able to adapt itself to the offspring of new tools and internal and external changes, especially those derived from the arising of new requirements or availability of new information. It must also foresee the great growth that this kind of systems usually undergo.

One of the main functions of the project director consists in setting the developing framework, evaluating the chosen solution, especially with regard to the availability of technology and data needed. That is how the implementation of solutions for which there is neither a technological nor organizational background is avoided [Gardner, 1998].

The setting of an architecture for the project, where the term architecture is conceived as “a set of rules or structures providing a framework for the overall design of a system or product”, brings out a number of advantages, namely:

- Improvement regarding information consistency, and subsequently, decisions taken from it.
- Enhancing communication with the management, as it is easier to show which are the purposes intended with the system, the fitting and workout in the rest of the corporate systems, etc.
- It enables the settling of an overall view for developers and technicians involved in the project, so that they can have knowledge of the reason for their work, and the way they are contributing to reach a determinate goal
- It helps the sorting out of necessary resources and tools, as it clarifies the different components and their correlative functions. Both hardware and software are included here.
- It reduces further maintenance costs, as it is easier to see the influence of changes on the whole.

We can conclude that the definition of a framework or architecture, on which both the data warehouse and the applications will be built, is one of the key factors for the success of the project [Kimball, 1996], both during its conception and development, and throughout its active life.

2.2 Data Warehouse Framework Components

Decomposing the data warehouse architecture results in three kinds of units:

- Storage elements, intended to house system information.
- Data handling procedures, also known as handling services.
- Human factor, consisting of final users and technical staff involved in the project

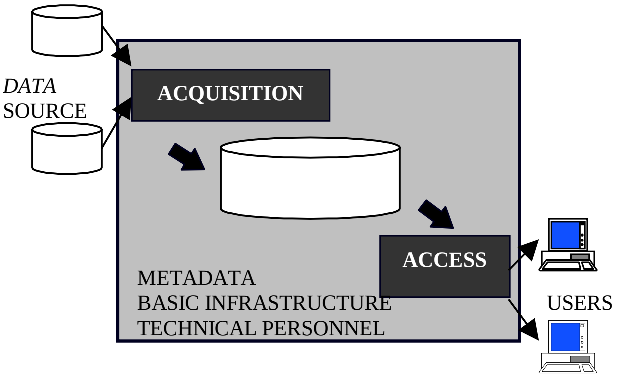

Figure 2: The data warehouse framework

Figure 2 shows the natural decomposing of a data warehouse, resulting in eight main components, each one embodied in any of the former categories, as follows: SOURCE DATA, ACQUISITION, STORAGE, ACCESS, USERS, TECHNICAL PERSONNEL, BASIC INFRASTRUCTURE AND METADATA. The last two are global encompassing elements and stand on the whole environment.

This is the classic schema for data warehouses, although it is possible to choose a simpler schema or even a more sophisticated one, as put forward in [Poe, 1996] for 2nd generation warehouses.

Although the most important piece of the framework is the warehouse itself, we must not neglect the rest of components, as the system would be worthless without them. Although this global view is conceptually simple, it will serve as a reference model for the enumeration of the functions regarding each of these elements, as well as for going deeply into the internal structure and problems related to each of these subsystems

On subsequent issues, each of these elements and processes will be developed, as well as the control objectives for each one.

2.2.1 Source data

They form the feeding for the warehouse and, at least in the classical schema, they are the only way for information input. They comprise all the information handled by the corporation's applications containing useful data for the warehouse: these are the named operational systems. Support for this information can be very diverse: data can be on traditional files (sequential or indexed), files with a determined owner format (ISAM o VSAM), or stored under a DBMS, usually relational or object-oriented, although hierarchical or net supported DBMS can still be found.

Special mention should be made of corporations provided with ERP (Enterprise Resource Planning) systems for any of their functional areas (Accounting, Human Resources, Invoicing, ...). These systems are subject to the problem that they do not show the customer the exact data model they are based on. Moreover, the number of boards is usually high and it has to be decided which of them contain the information we are interested in, and their relationships to the rest of information.

<u>Control Objectives</u>:

- The data warehouse should take into account the diverse types of information available (text, images, fax, video, sounds, ...) as well as information coming from external sources.
- Source data, both of internal and external origin, must be completely documented, including the data model in which they are based. If these models do not exist, they must be obtained, either manually or automatically, by means of specific reverse engineering tools. This will be necessary to commit

the data-mapping phase in the life-cycle.

- Source data security requirements must be documented, so that this security level can be ensured in the warehouse.

2.2.2 The acquisition process

Set of processes through which information is extracted from their original systems (operational systems, external systems, etc.), and after a number of transformation, integration and depression processes, is finally loaded into the warehouse, conforming to its data model.

Next, and following the scheme shown in figure 3, sub-processes in which *data acquisition* can be decomposed into, will be developed.

In order to enhance comprehension of these processes, it is convenient to consider that there is a temporal storing area, which will be used as a supporting element for every intermediate process between extraction and definitive loading. This area will neither be perceptible, nor consequently accessible to final users.

These sub-processes are described below:

2.2.2.1 Extraction

As soon as the location of necessary data has been identified, we must proceed to read and unload them, going for this purpose to the systems that store them. This process can take up to 60 % of the total time [Kimball et al., 1988]. It will be more difficult for old systems, especially in the case of mainframes, and also when extraction is made from different systems or database managers.

Another item worth considering is whether extraction should be made either totally or increasingly. In the later case, only information modified since the last extraction is extracted.

Eventually, if the destination for extracted information is a system different to the original, consideration or considerations can be made about compressing at the same time as extracting, in order to reduce intercourse with the destination system.

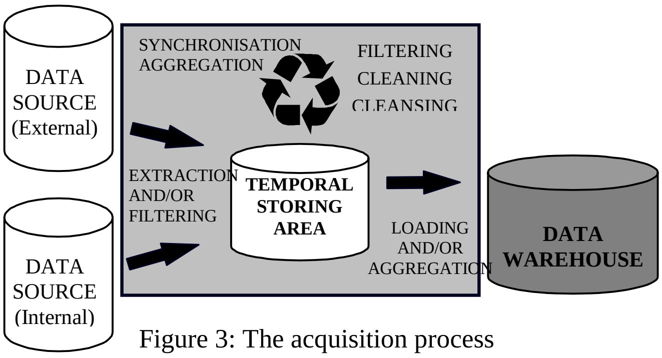

Figure 3: The acquisition process

2.2.2.2 Filtering

It consists of selecting, according to specific criteria, only part of the information from source systems. Among most usual criteria are temporal span, geographic area, etc. In many cases, this process can be performed simultaneously to extraction.

2.2.2.3 Cleaning and cleansing

They are intended to eliminate inaccurate or inconsistent information, by means of a number of checking and/or correcting operations. This process is very complex, and despite having specific tools, requires a great effort. Some aspects that should be checked are, for instance:

- Referential integrity checking of related entities.
- Data de-normalization if the data model requires it.
- Data format unification: its objective is that data corresponding to the same entity in the data warehouse have compatible data types and internal representations. Transformations will depend on the kind of data: numeric, date, alphanumeric, etc.
- Null value treatment: if source systems do not support null values, as it happens with old DBMS or traditional files, they often use a fictitious value to represent it. Representation among different systems must be unified.

2.2.2.4 Synchronization

Mechanism intended to give temporal reference to information coming from different systems. As operational systems seldom deal with historic values while the data warehouse does, temporal information must be obtained when it is not available. Information coming from different sources should be synchronized, so that it fits in consistent temporal intervals. Temporal reference can be associated to a whole table, to a record, or to each field. Usually it is related to each record, as doing it to each field supposes a considerable overload.

2.2.2.5 Aggregation

It is usual to record onto the data warehouse previously calculated information so that further access is faster [Harinarayan et al., 1996]. This process can be made with tools or either program it specifically. It can be done as an isolated process or simultaneously, with the conclusive loading in the data warehouse.

2.2.2.6 Loading

It consists of inserting the information inside the data warehouse once processed, following the data model previously fixed. Often, it implies creation or regeneration of every index. Several systems are usually involved and definition and planning of the different tasks must be clearly specified.

The receptor DBMS itself will usually possess some loading tool to ease this task. If it is not the case, some auxiliary tool may be used.

<u>*Control Objectives*</u>:

- Whenever possible, specialized tools will be used for every data acquisition process. In any case, the decision must be justified according to cost, necessary learning, further use in other projects, etc.

- Whenever a temporary loading area is used, its design should be as similar as possible as the definitive one, so that the loading process takes as few operations as possible.
- Data extraction must be accompanied by temporal information that allows further synchronization to other data sources.
- Procedures to check the operations involved in the data acquisition process should be settled. Therefore, each process should record control information about its work, including the number of records processed, sums of the main numeric columns, etc. Following each data loading, being either total or increasing, data warehouse information will be checked out and confronted to control information recorded during the acquisition process.
- Extracted data, once cleaned and cleansed, should conform to the minimum quality level settled in the data modeling stage.
- In every information intercourse between different systems, especially when public networks are used, security requirements settled in the data warehouse design, as well as those from source systems must be fulfilled. Procedures settled at this stage will complement the data warehouse security plan.
- Every process in the data acquisition, especially extraction and loading must be optimized in order to fulfil set time requirements.
- Aggregation and denormalization should be justified by a high access frequency or when the calculation time is too high to be executed in real time.
- A method to guarantee data coherence of aggregates should be settled.

2.2.3 Storage

This is the key component of the data warehouse, the place where all the data stand ready for their ulterior exploitation. Storage must be considered independently to further utilization made out of it, from both the users point of view and the applications point of view.

Thus it constitutes the framework nucleus and the basic part on which all of the other components lie. The final objective is to provide a transversal view to the organization, where different concepts become semantically unified.

In a simple data warehouse scheme, this is directly accessed by users, usually using specialized tools or ordering applications. However, the standard is a non-monolithic warehouse, but something capable to decompose, as shown in figure 4, into a series of elements and processes that relate them.

Among intervening processes the following stand out:

2.2.3.1 Replication

Through this process, the whole data warehouse is duplicated onto another system, sometimes for security reasons [Inmon, 1996b], i.e, ensuring disposability in case of system failure, and other times because of accessibility, in order to allow distant users to access the data warehouse, although the latter argument is becoming increasingly futile due to the evolution of high speed networks [Gª TOMAS et al., 1997].

2.2.3.2 Physical fragmentation

With this process, which is equally optional, the data warehouse is decomposed into sub- aggregates, not necessarily disjoint, that provide a partial view to the organization. This process is necessary to give access to the warehouse to users whose decision range is restricted to a part of the organization, but featuring the advantages of integration and access to quality information provided by the data warehouse. Each of these elements is named "data mart".

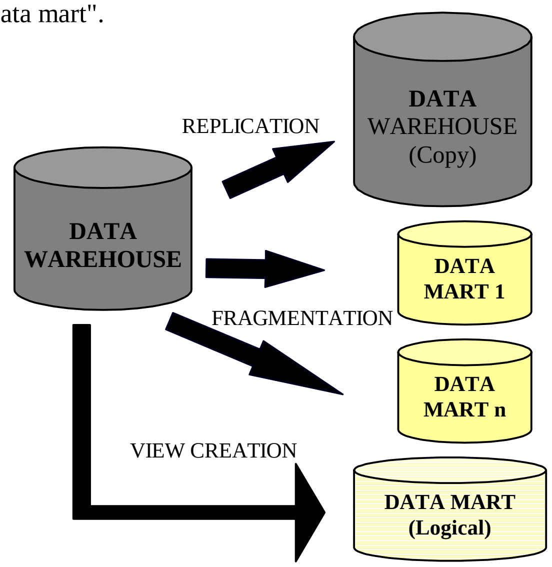

Figure 4: The storing area and the related processes

2.2.3.3 Logical fragmentation

This process consists in creating a logical view to the data warehouse. The effect is the same as with physical fragmentation, but without creating a new storage area or data mart. It has the disadvantage that when access is made through it, operations to show views must be processed in real time, increasing the response time.

A *Datamart* can be defined as a subject-oriented data base, disposable for users, for decentralized decision making, with lower range than that of the data warehouse. The amount of data marts will depend on the diversity of users.

There are authors, such as Kimball, who consider that users must always access the data warehouse through what he denominates presentation servers, which actually are data marts, and never directly. This matter is of no importance regarding the definition of the framework, as the elements are still the same. A single user could access more than one data mart to make composite queries, provided that they are designed using similar or compatible dimensions.

There are two main commercial options available for storing: On the one hand relational DBMS (Informix, Oracle, DB2, Sybase, ...) which often have a parallel variant, able to discompose queries into several processes that are executed separately, and on the other hand, multidimensional DBMS (i.e. Oracle Express, etc.)

Although multidimensional DBMS have the advantage of having aggregate information, achieving very good response times, they lack their limitations regarding capacity, which increases when taking into account that storing calculated information takes more space than strictly necessary. An added problem is that it is uneasy to store non-dimensional information, and, furthermore, the change in the data model is difficult once in use when information is loaded. All of these points make this option hardly viable for data warehouses nowadays, as they usually need a big capacity, and moreover, they have a very high growing rate. However, multidimensional DBMS may become an interesting option for data marts, as these have a smaller size, and they are the usual way of access for users.

The data warehouse is usually supported by a relational DBMS, which is the only type that supports the storing capacity required nowadays, supporting frequently as many as some terabytes. Problems associated to these systems are found mainly when the data model is complex and queries are made affecting many voluminous tables. Even though, when physical design is to be done, databases can be denormalized, and so aggregates can be calculated in order to reduce response times [Mumick et al., 1997]. In the case of a parallel version, this problem is partially solved, as querying times decrease.

The choice for relational or multidimensional DBMS is one of the key choices and one of the commercial battles in the area.

Control Objectives:

- The periodicity for replication processes, when these exist, must fit the storage updating rate, and in any case, the system failure risk and the organization level of dependency. Although all information can be retrieved, the maximum time allowed to be out of order points the choice to be made. This process must be included and documented in the system security plan.
- Data marts must never be considered an alternative to the data warehouse, but something that complements it. In initially committed in a partial form cases, with the construction of a data mart, the global and transversal view to the data warehouse should not be neglected, and specially, independent data marts construction should be avoided.
- Data marts will be subjected to equal security restrictions as the source data.
- The choice for the Data Base Management System (DBMS) must be carried out according to access time restrictions and storage volume, taking into account especially the predicted growing rate. Classic factors are not to be neglected: interconnection capability to other products, conformity with standards, provider's solvency, product sales share, price, etc.
- Procedures to monitor Data Base Management Systems should be settled in order to guarantee the system performance in the operation of the data warehouse.

2.2.4 Access

This item is constituted by the collection of processes intended to capture and exploit data warehouse information, so that users can be provided with the information they need for decision making. Access is done through a combination of standard tools and specifically designed programs.

The user should perceive that information is picked from a single point. For this reason, there must be a layer hiding the existence of the data warehouse and eventual data marts, unifying views. Access can be discomposed into several sub-processes:

- Query definition. There must be a mechanism for the user to express his/her need, usually on the basis of the existing information inventory. Thus, it is necessary that metadata exist and are reliable.
- Search for information. On the basis of the user's need, required information must be found, either in the data warehouse, or in one or more data marts.
- Information processing. Information has to be treated usually before it is shown, including unions, operations, filtering, etc.
- Showing information. The result of searching and processing must be supplied to the user that required it in a way that allows easy interpretation, and in a framework as user-friendly as possible.

The nucleus for this item is the management of queries, that is, the work done since an information requirement is made until this is executed and requested data are supplied. Queries simplification whenever possible, access to multiple systems depending on metadata, detection of already performed aggregates, planning of queries, etc. should be included here.

As we show in figure 5, there are two ways to access information: first one, which can be considered "traditional", according to which, given an initial hypothesis, the user tries to verify if it is accurate or not. For this purpose both fixed report generators and even parameter programmable ones can be used, as well as totally "ad hoc" report generators in which the user can define queries according to his/her need for information on each occasion. However, simple spreadsheets, specifically developed applications or EIS/DDS packages can also be included. Among access tools, a special mention should be made to OLAP tools [Ho et al., 1997]. These tools provide the user with a multidimensional view to the business, thus permitting to view data in a desired detailing level and analyzing interesting dimensions each time [Chaudhuri et al., 1997].

The second way to access, and thus exploit data warehouse information is the so-called *data mining*. Data mining is the group of processes and technologies intended to analyze data warehouse information in order to obtain knowledge not easy to perceive at first sight. This would include trend attainment, data correlation, information segmentation, etc. The innovation implied in this process lies in the fact that it is the system itself that makes choices and tries to discover valuable information through data associations [Berry et al., 1997]. Data mining can be included inside the process known as "Knowledge discovering in data bases" [Piatetsky-Shapiro and Frawley, 1991], more familiar as KDD (Knowledge Discovery in Databases). The techniques most frequently used in this process are statistics, artificial intelligence, decision trees, clustering, and, of course, they should be performed with specialized tools. As it can be seen in figure 5, the result of data mining activities can generate information that is stored again in the data warehouse.

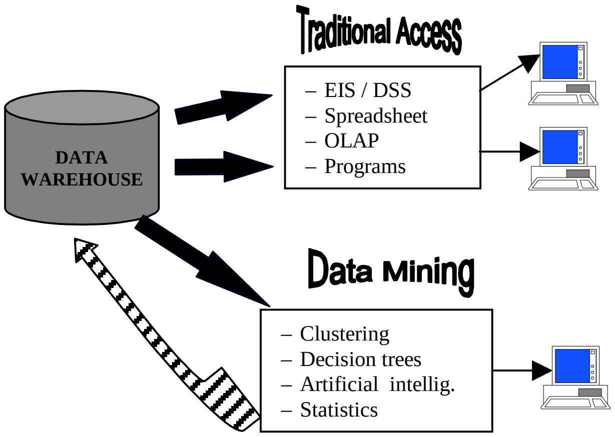

Figure 5: The access to the information

For these purposes, as well as for the rest of the processes, specialized tools capable of solving interaction with users and perfectly accessing to implied DBMS are frequently used.

Regarding the way of interaction, usually is the user who requires information when s/he needs it. However, there is a new way to access, called "data push", in which is the system itself that, when relevant changes according to a number of previously programmed parameters are detected, performs pertinent transformations and reports this circumstance and associated data to the user.

Control Objectives

- Transactional applications lack regarding reports should not be tried to be solved through the data warehouse. Even in the case of simple reports, these must be oriented to decision-making.
- Reports and queries must be parameter-programmable, with capability to be run in an iterative way as a function of parameters, to be planned, to be distributed through a diversity of means, etc.
- The user must perceive only one view to information, independent from its physical location and origin.
- The accessing environment must be as simple and intuitive as possible, and be completely guided by metadata.
- System operation should be monitored, specially the queries and tables most required, sizes of tables and indexes, frequency of users' use, etc.. All this information will be employed to improve the system performance, creating or deleting indexes, aggregates, etc., as well as to justify the investment and to plan the future of the system (new processors, disks, networks, etc.).
- Accessing should fulfil every system security requirement, with special care for authenticating and authorizing processes; that is, it should be checked that the user actually is who s/he says s/he is, and that s/he is allowed in every case for a specific operation.

2.5.2 Users

They are business analysts, managers, etc., who are willing to improve their decision taking, basing on the information they are going to extract from the data warehouse. The data warehouse is intended for them, and they are thus the final receivers of all the processes, tools and applications described in this framework. Among them, we can discriminate managers with a responsibility over the complete organization, who access the data warehouse as a whole, and managers who, owing to their restricted management area, cannot access the whole data warehouse. In this latter case, they will use one or more of the data marts, if these do exist, and if not, the security policy will restrict their access to the authorized part of the general warehouse. Optionally, and though it is not the main purpose for the system, access to the whole or part of the data warehouse can be given to non-manager users, or even non-organization users.

Control Objectives:

- The users must take an important part in the definition of the system requirements, being part of the project from its conception and approval to its ending.
- An individual control must be carried out on the evolution of the data warehouse use by each user once the system has been implanted and in use.
- Training needs for users should be identified and documented basing on both the system functionality

and the different users' skills.

2.2.6 Technical personnel

Although they are not usually regarded as a part of the framework, they are indispensable for coordinating the rest of elements, managing the project in order to fulfil their objectives.

All personnel needed for the running and further operation of the data warehouse are included here, marking the necessary peculiarities and abilities required by this kind of projects. Diverse profiles to be taken into account are the following:

2.2.6.1 Project director

S/he is in charge of the control and co-ordination of the project. S/he must dominate management techniques that can be applied to other kind of projects, although he will have to know the technologies and peculiarities of the data warehouses, especially the life-cycle, existing tools on the market, etc [Inmon et al., 1997].

2.2.6.2 Analysts

Their function, as it happens with other kinds of projects, will be the definition of the system requirements and design of solutions. The main difference lies in the strong orientation towards data implied in these systems, in which processes are of less importance. It will deal mainly with defining the data model, from a logical or conceptual point of view, but not forgetting the physical issues related to dealing with different systems: relational, multidimensional, parallel relational, etc. Furthermore, classic modeling techniques cannot be applied as a whole, as normalization should not be exhaustive and in many cases different schemes are pursued (star or snowflake). The project range is fairly wide, that is the reason why a certain experience in big projects and a good abstraction capability are needed to de able to manage all of these functions.

2.2.6.3 Database administrators

Their function consists of the physical design for each of the DBMS involved, and their optimization later on, as well as setting security procedures as a function of the requirements. Usually several DBMS, operating systems and hardware are frequently used, and so it is necessary know each of these environments in depth [Labio et al., 1997].

2.2.6.4 Application developers

Their objectives within the project consists of carrying through the programming and implementation of the software involved in all defined processes, and specially:

- Configuring and programming access tools, usually based on the client/server philosophy.
- Programming of the processes of extraction, cleaning, cleansing and loading, either through standard programming language programs or using specialized tools on these tasks.
- Programming of the processes of data mart fragmentation. In the same way, it can be carried out through any specialized tool or "ad hoc" programs.
- Data mining process programming or configuration of tools in use.

2.2.6.5 Network managers

Their function is similar to that in other projects, although as huge data volumes are sent through the network, optimization plays an even more important role.

2.2.6.6 Hardware technicians

Beside usual knowledge, they should be specialized in specific systems, usually multiprocessor systems, high performance systems, etc.

2.2.6.7 Operators

Among their functions are, besides typical start/stop operations in different equipments, communication systems, DBMS, etc., security copies and performance controls, taking charge of the execution on a previously fixed periodic basis, of extracting, filtering, cleansing, cleaning, and latter loading into the data warehouse processes. For this purpose they need specific complementary formation, as the number of tools to be employed is usually high, and most of them have high/medium complexity.

Control Objectives:

- Depending on both the functionality of applications and the architecture and infrastructure on which the project is supported, a project team must be defined, which is capable to carry through the different kinds of tasks (requirement analysis, design, tool/system configuration, programming, operation, etc...)
- The whole technical team, or at least most of it, must be previously experienced in this kind of projects.
- The project team should have experience with the selected tools. Otherwise, the necessary training should be included inside the plan.

2.2.7 Basic infrastructure

Except for the users and the technical personnel, the rest of elements composing the data warehouse environment require some hardware to lean on in order to reach their purpose, that is, storing or processing information. Moreover, it is necessary to provide means to interconnect the different systems involved. This basic infrastructure of hardware and communication systems, added to base software, does not differ greatly from that needed for other kind of systems, although it has some peculiarities.

Very frequently, the data warehouse infrastructure is shared with other organizations systems, resulting in a slightly diffuse border of the data warehouse. For a correct definition of the infrastructure, the following factors must be considered:

- Data volume to be handled.

- Data volatility, depending on actualization frequency.
- Number of users and their foreseen activity, as well as their location.
- Number of business processes.
- Kind of use, bearing in mind if standard reports will be used, which can be previously optimized, or “ad hoc” queries.
- Availability of certain software products, that are only available for some platforms.
- Availability of technicians with qualification on the environment.
- Economic resources available.

Hardware includes every platform for each data warehouse, application server and even users’ computers. Generally, the most accepted option is the open systems use, as productivity, tool availability and interconnection capability are by far superior to mainframes.

When data volumes are very large, advanced architectures are necessary to respond according to time requirements. It is usual to find multiple processor systems, with subsequent associated administration complexity. Most common options are SMP and MMP.

In any case, disk velocity and time for memory access are decisive, as the volume of data handled in each inquiry uses to be very large. Memory amount must be as big as possible to reasonably avoid disk access, being 64 bits optimum.

The requirements for the final users system, generally known as desktop, will depend on the applications that are to be used, as it can range from a simple web browser accessing a web server, to a powerful data mining tool. This choice will be made at the relevant time in the life cycle and will thus determine the kind of system required.

In the item on storage several options for the Data Base Management Systems (DBMS) have already been discussed.

In order to connect different systems, available LAN/WAN networks are generally used, although research should be done on regarding if bandwidth is enough for this new use. Special attention must be paid to supporting databases through ODBC, JDBC, etc.

Control Objectives:

- The infrastructure should be stable throughout the project.
- The election of hardware and software tools should be carried out according to objective criteria: supplier's solvency, price, easy of administration, agreement with the standards of the organization, easy interconnection with other tools, ...
- The different tools elected should be able to exchange metadata.
- The infrastructure should contemplate at least: hardware, tools of extraction, cleaning, cleansing, transformation and load, middleware, metadata tools, DBMS (RDBMS and/or ROLAP DB and/or MOLAP DB), administration and security tools and access tools for users.
- All the infrastructure elements should have enough growth capacity to support the future system load with acceptable response times. It will keep especially in mind the number of users and usually very high data volumes. The suppliers' promises about the tools growth capacity should be proven.
- A specific security plan should be developed in order to guarantee the continuity of operation and the availability of the system. It should address both logical and physical security, as well as a contingency plan for emergency situations and backup and recovery procedures. It will deserve a special consideration in those cases where the warehouse is related to the Internet environment [Inmon, 1996b].

2.2.8 Metadata

They are the base and the support for every data acquisition and accessing process. Traditionally, metadata are defined as "data about data". As it has been seen through out this paper, the data warehouse environment is complex, and subsequently, so will be metadata associated to it. It is necessary to record metadata on:

- Organization: business rules, functions, responsibilities.
- Source systems: specifications for accessing information, database documentation, information description, legal limits, security access, owners, actualization frequency, ...
- Processes: extraction, loading, cleansing, filtering, temporization, etc., including rules to follow in each of them.
- Data warehouse and data marts: related data model, tables, columns, indexes, location and information security, etc.
- Queries and reports, including description of the information they provide and security associated to them.

Availability of metadata about all of these elements has a number of advantages:

- They isolate the data warehouse from changes in operational systems.
- They contribute to provide a unified view to the data warehouse. This is a logical view, as each tool includes its own metadata management, and there is not an interface that allows interaction between tools in this way. So what should be done is to keep this documentation currently and centralized.
- They serve as a nucleus for every transformation

affecting information, from its origin to the data warehouse and then to users.

Metadata can play either an active role, if they are useful to lead some processes, such as creating a table with a procedure, or a passive one, as simple documentation.

The concept of metadata is, as we have seen, very comprehensive, as it comprises the whole project. The right compilation and maintenance of metadata, which is the equivalent to documentation in traditional systems, is one of the key elements of the project [Daruwala et al., 1995].

Control Objectives:

- A corporate dictionary should exist to maintain metadata. It should be updated and include at least: definition of views, algorithms to aggregate data, data mappings, users, ... Metadata are the core of a data warehouse and should be managed accordingly.
- Metadata management must be automated by means of a specialized tool, and must never be embedded in programs or procedures.
- If metadata co-exist in different tools, these must be able to share this information and show it to the user in a conjunct way.
- Metadata must be dynamic and evolve through the project and the system lifetime, being actualized every time that any changes that affect them occur.

3 The audit process

The audit process consists of checking the established control system. The general objectives of auditing [ISACF, 1998b] are:

- To provide management with reasonable assurance that control objectives are being fulfilled.
- To substantiate the risks caused by a control weakness.
- To advise management on corrective actions.

The methodology we propose in order to carry out this process is based on assessment of risks. Risk is defined as a chance for injury, damage or loss due to using data inappropriately [Curtis and Joshi, 1997]. Starting from the inherent risks that threaten the project, the control objectives that minimize those menaces will be settled. The objectives have been defined in the previous section, although they should not be interpreted in a strict way and should be adapted by the auditor to each project and organization.

The goal of control objectives is to have information that satisfies the business requirements. This goal is achieved through seven criteria [ISACFa, 1998]:

Effectiveness: the information has to be relevant and pertinent to the business process, as well as delivered in a timely, correct, consistent and usable manner.

Efficiency: the provision of information through the optimal use of resources.

Confidentiality: the protection of sensitive information from unauthorized disclosure.

Integrity: relates to the accuracy and completeness of information, as well as to its validity in accordance with business expectations.

Availability: the information has to be available when required by the business process. It also involves the safeguarding of necessary resources and associated capabilities.

Compliance: deals with complying with laws, regulations, and contractual arrangements the business process is subjected to.

Reliability: the provision of appropriate information for management to operate the entity.

Given a data warehouse project, a specific audit plan should be drawn to achieve the described audit goals. It is necessary to take into account both the drawn control objectives and the resources available for the audit process (including people, budget, tools, ...), so that the plan is realistic and workable [Piattini and Del Peso, 1998]. The reached detail level in the audit process depends on the available resources to execute it.

To accomplish the audit we propose the four stages shown in figure 6:

AUDIT STAGES

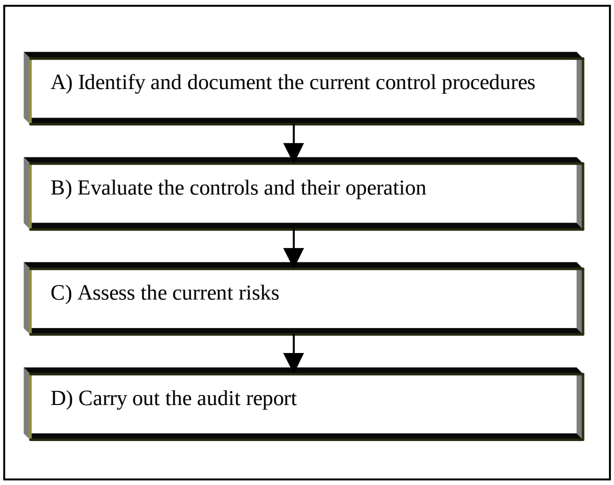
A) Identify and document the current control procedures
B) Evaluate the controls and their operation
C) Assess the current risks
D) Carry out the audit report

Figure 6: The audit process

A) Identify and document the current control procedures.

The purpose of this stage is the comprehension of the business; including its requirements, risks and policies, the organization structure, the different roles and positions, ... Through this activity, the auditor will identify and document the detected control measures. This is achieved by both interviewing the adequate staff and reviewing the related documentation.

B) Evaluate the controls and their operation.

This phase tries to settle whether the current measures of control are effective in order to get the established control objectives. The result will be the degree to which the control objectives are met. It is necessary to check if:

- There are enough controls where it is necessary.
- Controls are adequate to assure the business against potential risks.
- Controls work in a continuous way along the project. The auditor has to take into account this time consideration, because the whole project is under revision and not only the delivered products.
- Controls operation is documented appropriately, including the task checked, the result, the person responsible of the revision, date and time, comments, ...
- There is an adequate workflow to report the detected problems to people responsible for the tested process.

To perform this phase it is necessary to obtain direct or indirect evidences about the issues enumerated above. The auditor has to review the documentation about controls and their operation, as well as to interview the appropriate people in the project team

C) Assess current risks.

Once the business requirements and all the information about existent controls and their operation, the auditor has to obtain a measure of the current risk. This measure should be based on the detected control weaknesses and their related threats. For each menace the probability of occurrence should be settled as well as the damages it could cause to the organization and its impact.

This assessment task is very difficult and to be performed in a right way it is necessary a great experience and capacity of analysis.

D) Carry out the audit report.

The result of the audit process ought to be expressed in a written report. It should include the objectives of the audit, the period of coverage and the nature and extent of the audit work performed, as well as the related conclusions and recommendations [Moeller, 1989]. The addressees of the report will be managers of the affected departments.

4 Conclusions and future works

Data warehouses are a field with an important present, as it is proved by its witnesses in the market, as well as with an important future, as it is shown by the high number of current lines of research.

Taking into account that data should be considered as a corporate resource, subject to the usual control and audit tasks, in this paper we have put forward:

- A definition of a formal framework for data warehouse projects.
- A number of control objectives for each component of this framework. They are based on COBIT philosophy and pretend to minimize the related risks.
- A methodology for the audit process.

Basing on the contributions made in this paper, we suggest the following future research guidelines:

- The refinement of the proposed control objectives through their application in real projects.
- The proposal of a set of metrics for the different control objectives.
- The development of a tool in order to make easier the audit process, especially with regard to evaluation of controls, assessment of risks and reporting of results.

All of these guidelines are currently being developed in the framework of the following project: a data warehouse for the Regional Government of Murcia (Spain) to manage data about economic in the last twenty years. The system pretends to improve political decisions about unemployment, industry, population ..., achieving an optimal distribution of budgets in next years.

References

BERRY, M.J.A. and GORDON, L. (1997).
"Data Mining techniques". John Wiley & Sons, New York.

CHAUDHURI, S. and DAYAL, U. (1997)
"An Overview of Data Warehousing and OLAP Technology". *ACM SIGMOD Record* 26 (1), pp. 65-74.

CURTIS, MARY and JOSHI, KAILASH (1997)
"Internal Control Issues for Data Warehousing". IS Audit & Control Journal, Volume IV.

DARUWALA, A., GOH, C., HOFMEISTER, S., MADNICK, S. and SIEGEL, M. (1995)
"The Context Interchange Network Prototype". Proceedings of the Sixth IFIP TC-2 Conference on Data Semantic (DS-6).

GARDNER, S. (1998)
Communications of the ACM, vol. 41, Nº 9, September, pp 52-60.

Gª TOMAS, J., FERRANDO, S., and PIATTINI, M. (1997)
"Redes de Alta Velocidad". Ra-ma, Madrid.

HALEY, B., and WATSON, H. (1998)
"Managerial Considerations". Comm. of the ACM. September, pp 32-37.

HARINARAYAN, V., RAJARAMAN, A. and ULLMAN, J. D. (1996)
"Implementing Data Cubes Efficiently". Proc. of the 1996 ACM SIGMOD International Conference on Management

of Data, Jagadish, H. V. and Mumick, I. S. (eds.), pp. 205-216.

HO, C-T., AGRAWAL, R., MEGIDDO, N. and SRIKANT, R. (1997)
"Range queries in OLAP data cubes". Proc. of the 1997 ACM SIGMOD International Conference on Management of Data. Peckman, J.M. (ed.), pp. 73-88.

INMON, W. (1996a)
"Building the Data Warehouse". 2nd edition. John Wiley & Sons.

INMON, W. (1996b)
"Security in the Data Warehouse/Internet Environment". IS Audit & Control Journal, vol. IV, pp. 8-11.

INMON, W., WELCH, J.D., and GLASSEY, K. (1997)
"Managing the Data Warehouse". John Wiley & Sons.

ISACF (1998a).
"COBIT Audit guidelines". Information Systems Audit and Control Foundation. Rolling Meadows. IL, USA.

ISACF (1998b).
"Control Objectives for Information and Related Technology". Information Systems Audit and Control Foundation. Rolling Meadows. IL, USA.

KELLY, S. (1997)
"Data Warehousing in Action". John Wiley & Sons.

KIMBALL, R. (1996)
"Practical Techniques for Building Dimensional Data Warehouses". John Wiley & Sons.

KIMBALL, R., REEVES, L., ROSS, M., and THORNTHWAITE, W. (1998)
"The Data Warehouse Lifecycle toolkit: Expert Methods for Designing, Developing, and Deploying Data Warehouses". John Wiley & Sons.

LABIO, W., QUASS, D. and ADELBERG, B. (1997)
"Physical Database Design for Data Warehouses". *Thirteen International Conference on Data Engineering,* IEEE Computer Society, Birmingham, UK, pp. 277-288.

MOELLER, R. (1989)
"Computer audit, control, and security". John Wiley & Sons.

MUMICK, I. S., QUASS, D. and MUMICK, B.S. (1997)
"Maintenance of Data Cubes and Summary Tables in a Warehouse". Proc. of the 1997 ACM SIGMOD International Conference on Management of Data. Peckman, J.M. (ed.), pp. 100-111.

PIATETSKY-SHAPIRO, G. and FRAWLEY, W.J. (1991)
"Knowledge Discovery in Databases". AAAI Press/The MIT Press. Menlo Park. California.

PIATTINI, M. and DEL PESO, E. (1998)
"Auditoría Informática: un enfoque práctico". Ra-ma (Madrid, Spain).

POE, V. (1996)
"Building a Data Warehouse for Decision Support". Prentice Hall.

PREECE, J., ROGERS, Y., and SHARP, H. (1994)
"Human Computer Interaction". Addison-Wesley, Wokingham (GB).

RODERO, J.A., PIATTINI, M.G. (1998)
"Auditing Data Warehouses through the life-cycle". Proc. of the BDATOS'98 conference, Buenos Aires, Argentina, pp. 9-20.

SPRAGUE, R., and WATSON, H. (1996)
"Decision Support for Management". Prentice-Hall, New York.

VANDENBOSCH, B., and HUFF, S. (1992)
"Executive support systems and management learning". Business quarterly, University of Western Ontario, Autumn, pp. 33-38.

WATSON, H., RAINER, R., KELLY, and HOUDESHEL, G. (1992)
"Executive Information Systems: Emergence, Development, Impact". John Wiley & Sons, New York.