Data Warehousing: A Methodology for the Construction of Stores of Information

Leopoldo Galindo Soria[1] and Oscar Camacho Nieto [2]

[1] Sección de Estudios de Posgrado de la ESIME ZAC - IPN
Edificio 5, 2°. Piso, Área de Ingeniería de Sistemas.
Tel. 5- 729- 6000 Ext. 54805
lgalindos@yahoo.com.mx
[2] Center for Computing Research-IPN, Av. Juan de Dios Bátiz Esq. Othón de Mendizabal S/N, Col. Nueva Industrial Vallejo, 07738, México, D.F., México
oscarc@cic.ipn.mx

Abstract. In this document, appears a type of methodology for the construction of data stores (Warehouse Data, by its English denomination in and DW, by its abbreviations), and normally, are very big as far as storage capacity; with the objective to present/display it to the people interested in the new information technologies. The activities which appear are presented and explained, such as: identification and definition of needs, planning, strategy, identification of data sources, modeling of the data structure, design of Data Base, mapping, extraction, cleaning and transformation of data, loading and creation of applications in the DW, the tests and validation, as well as the total liberation. Also, they comment some possible techniques and tools for its construction.

1 Introduction

It is considered that the Information Technologies take companies and to their clients to obtain better services and therefore greater yields due to a better decision making based on its opportunity and pertinence and with this increase its position in the market. One of the IT that is arising not because of its applicability but because of its conceptualization and modelization is Data Warehousing.

That is, the process of obtaining, purifying, integrating and structuring the historical and present data of an organization in a warehouse or data base; based on this information for a better and possible real time making of decisions

Basically when a Warehouse Data Base is created we must take into account the following characteristics:

- Be oriented to any institution.
- Administrate large amounts of information.
- Integrate several versions of modeling schemes of the information of the institution that has been given in the course of the time.
- To Condense, integrate, associate and to present the majority of the prominent information of the business.

© *A. Gelbukh, C. Yáñez Márquez, O. Camacho Nieto (Eds.)*
Advances in Artificial Intelligence and Computer Science
Research on Computing Science 14, 2005, pp. 203-210

All these activities, are defiant tasks by the diversity of: the technologies of original storage, the recovery of the data or of the information, its filtered for its storage and common processes, its integration, its common structuring , and finally and not less important, its presentation in the adequate and most optimum form for the user that has to make decisions.

Considering the previous, as a follow up, appears a methodology in which the activities that are described, are considered necessary for the creation of data storage with the characteristics of Data Warehouse. In Figure 1, a scheme is shown with its activities and their possible interactions.

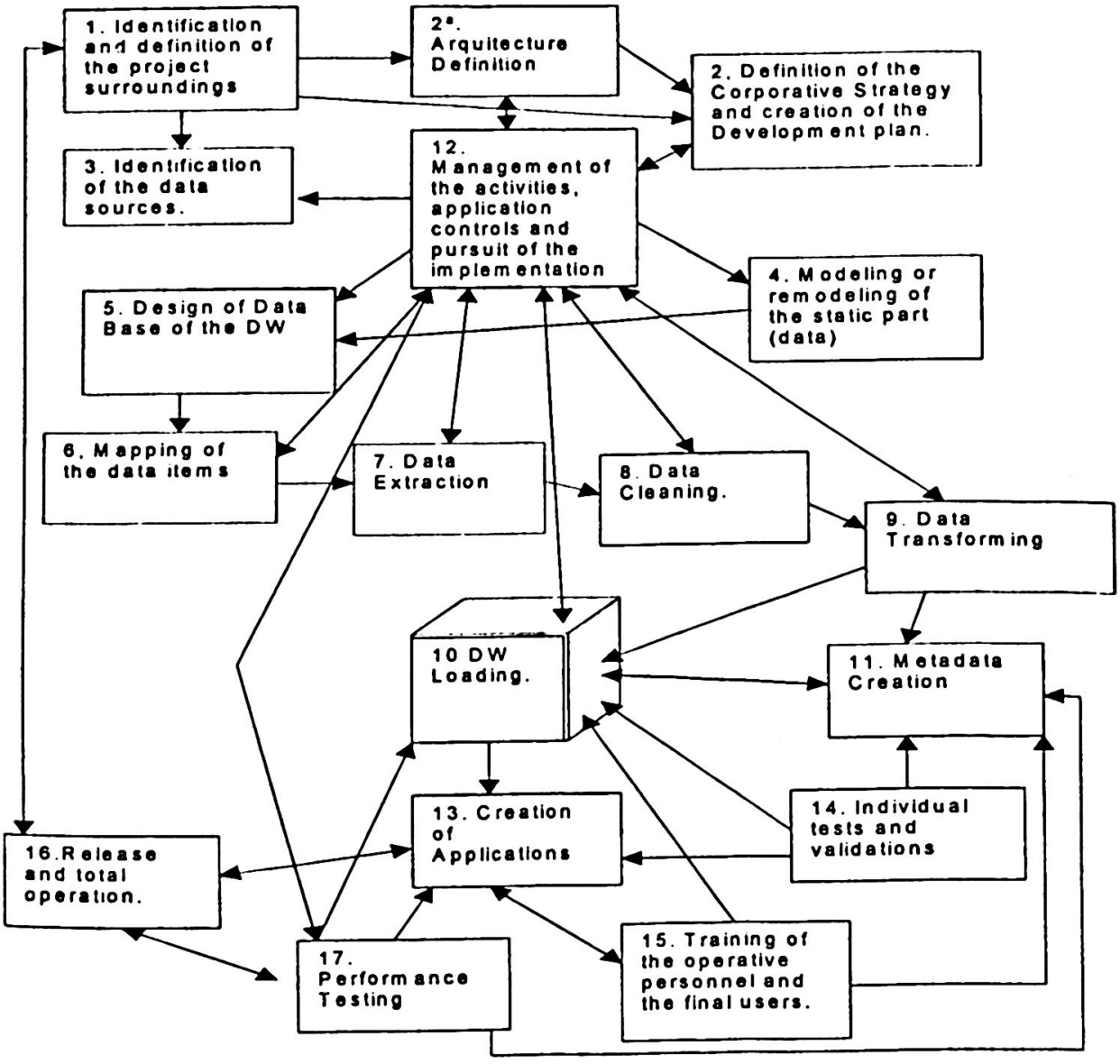

Figure 1. A scheme with its activities and their possible interactions.

2 Proposed Methodology

The following activities are the suggested activities to develop:

Activity 1: Identification and definition of the surroundings/environment of the project and compilation, of all types of documentation and information.

We must carry out the following activities:

Value the needs and the possibility of obtaining the necessary resources.

- Understand as best as possible the company and the involved areas.
- Identify the actors and key elements as well as its style. We must consider executives which have the knowledge to make decisions who can not answer/take it at the present stage.
- Design, apply and analyze the interviews as web as the questionnaires.
- To compile and review the existing documentation on the problem of the decision making and the operational systems that exist at the present time

It is very important to identify at this stage: Who? (Humans, what and/or systems) administer the processes of the business? Who? (Humans, what and/or other systems) administer the systems: control, production, computer science or others?
Who? (Humans what or and/or systems) administer the data?

Besides we have to get information about:

- Functions of the business, their factors of success or failure, their location, etc.
- Technology, hardware, networks and Internet, systems of relational databases or any other, access and development tools of data.
- Data: corporative model of data, data by area, distribution and physical definition
- Consulting, delivery times, quality, retention, history, metadata.
- Inherited "or present Applications, reports, systems and files", planning of future systems.
- Organization, origin of the resources for the project, structures of the roll area, Information systems and responsibility for the development of project and Information systems of Warehouse Data

For this, it is suggested to use the compilation techniques and classification of the information, as well as the interviews and it is due to obtain, the definition of the surroundings/environment of the project and a folder with all the documentation and obtained data.

Activity 2: Definition of the Corporative Strategy and Creation of the Development plan

It is necessary to identify and analyze in detail the following domains of the company:

- Proceses Domain.
- Data Domain.
- Information Systems Domain.
- Making Decisions Domain Support.
- Human Resources Domain.

And then, get done the following objectives:

- Identify the vision and the planning in the long term.
- Establish a reference frame for future developments.
- Establish consensuses on common objectives.
- Identify the key requirements of the creation of the DW infrastructure.
- Establish time of initial issue.
- Obtain the commitment of the direction

From this, the plan sets out, in which the following aspects have got to be included:

- Value the present situation.
- Identification of the reachment.
- Identification of tasks.
- Definition of resources.
- Allocation of resources.
- Definition of delivery times.
- Product Definition to deliver.
- Development of contingency plans.

Activity 3: Data sources identification.

It is required to make the following things:

- To identify the origin of the data:
- traditional Internal
- nontraditional Internal
- Image, video, text, voice or music, Internet, etc. External traditional or not.

To identify: where the data is originated, the tools to use to extract, to clean and to transform the information, what support platform is used, what data models are had, what type of data base systems are being used

To identify in detail how it will be the access to the information.

Activity 4: Modeled or remodeling of the static part (Data).

This is a fundamental activity for the later design with greater possible success of the data base of the DW.

It is necessary to identify or to redefine the data models of the systems of data bases from where the original information will be taken.

We must consider that the data modeling of the DW, will be different from one classic one, let's say of the relational type, in the following thing:

- Denormalized.
- We have as much detailed data as summarized or weighed.
- Oriented to, that is possible to make decisions on the basis of them.
- Must contain the time as a key
- Data derived from other data and strategists.
- May contain data adjustments.
- Organized around its use and stability

Activity 5: Design of the Data Base (Scheme Star) of Data Warehouse.

This activity is fundamental for the creation of a DW, since from it, it depend the possibility of an optimal physical storage of the data, its administration and its access. In addition, we must remember that at this point, we wish to construct "a bucket of information" or multidimensional data base (BDM), that is different from the flat tables of the relational model.

For this aim we model or design the "Scheme Star or Star Scheme", that structures in two types of denominated elements: dimension tables and the facts table, this last one integrates of all those of dimensions and will serve as a link among all the elements of the multidimensional table.

Activity 6: Mapping of the data items.

Here, two main activities are made:

- Problems Solving of: definition of concepts, homonyms, synonymous, etc., of the sources of original data, as well as, the possible fields or registries to derive that they can be added or accumulated, dimensional, balance, totals, by region, time, products, etc., and
- Create the consulting scenes and their possible equivalences, thus to determine if it is possible to obtain that data and to avoid in addition, possible redundancies in the storage of the data

Activity 7: Data Extraction.

Once perfectly located the data sources, we must consider the following thing for its extraction:

- Specific Elements to obtain: archives or specific parts of them, views or consultations, etc. This is very important, since it will be the physical base of the construction of the information "bucket".
- Define the activity to make with these data structures, as it can be its: replication, copy or if we will create access links, only at the moment of a possible virtual creation of the DW.
- Define and acquire or to create the tool of extraction of the information. It is possible to say that, there are many tools for this aim, but even, it's possible to use languages like C or Cobol or SQL, in a given time.

Activity 8: Cleaning of data.

Here, it is required to work in the main problem that exists when extracting the data of its base source: the integrity and quality of the obtained data. The Methods that are used, use defined processes or on the operations of the business thus to derive correct data from incorrect or incomplete data. In addition, they are due to solve for example, synonymous problems of when receiving the same information of diverse sources. Just doing this last one, it is a "gain" for the company

Activity 9: Data Transforming.

Once the data has been "extracted" and "cleaned", we must transform it for its later loading, that is to say, for example, that has different formats for the same data, as it can be "sales of the day" and for its integration it is necessary that they have a same format, type, mnemonic or labels, etc. Solving the previous thing in automatic form for any process of this type or resemblance, as it can be the interoperability of multiple systems of data bases, is indeed, one of the central problems of the investigation in data bases at the present time. This is clear, although many and diverse commercial systems exist that make this type of activities.

Activity 10: Load of Data Warehouse.

This is the most complex and difficult physical activity, which it is done in all the process and it requires that the entire corresponding infrastructure is ready. This is; the computer and network platforms, the data must be already: extracted, purified and transformed and ready or accessible, the computer system that makes the physical construction of the data "bucket", the trained personnel or enabled, ready to solve any eventuality. In extreme this activity is complex and demanding of all the resources of the project and in agreement with its magnitude can be necessary working several days (with its nights), in uninterrupted form to manage to conclude it.

Activity 11: Design and construction of the Metadata.

Now we must design and construct a suprasystem, in which the information is structured about the content of the information stored in the DW. This system is known as "Metadata", and identifies and specifies the structures of data in DW and how they are administered. In addition, it includes "triggers" of extractions, loads and updates. That is to say, it contains data about the data, such as: names and definitions of fields, "maps" of data, tables, indices, criteria of selection, and calculations of the derived data

The first stage of creating the metadata, is based on designing it for it, is required to model it, under a scheme for example, type (E-R), and from constructing it there and physically structuring it.

Activity 12: Settlement of the administrative processes, control and total implementation.

We must define and create throughout the entire project, administrative processes that allow managing:

- Processes: Controls, endorsements, updates, etc.
- The involved computer systems of: origin, destiny, support, etc.
- The diverse data and metadata.

As one can see, creating complex and integral systems, entails to have the systemic vision of not only solving the technical problem but, also all the inherent administrative process.

In the case of the DW, multiple and diverse computer tools for the support of all the process of management exist, but obviously they are expensive and plaintiffs of many resources.

At this point, a possible subject of investigation and the creation of a commercial product more competitive even exist than the present ones

Activity 13: Creation of applications of Data Warehouse.

Essentially, the DW does not have a goal without the use of tools that supports the users in the process of decision making, that in addition is designed with a direction on form and in style of being those people. Therefore, the friendless and the functionality of the system are their key elements. It is possible to say that the mentioned technology of construction of DW, has supported in a bigger way the creation of the "Decision Support System for the Decision Making", that presents/displays the information stored in data bases, with added values, such as: statistical groupings, with its respective analyses, in addition to all type of graphics, images, video and even now voice. Another type of systems including in these aids is the called Expert systems, which use concepts of the denominated Artificial intelligence, for "making decisions" or to orient in a wide form to the user in it, being based on denominated premises "inference rules or deductive" combined to other related concepts, such as, the "Data Mining". As it is observed, the presentation of the information to the decision maker of decisions, is a very complex and important activity, and involves the election or construction of the best tools to manage to obtain the best advantage of the "bucket of information".

Activity 14: Tests and total validation of Data Warehouse.

Now, an exhaustive process of test of each one of the elements is involved, since the extraction of the information source it is required up to the presentation of the results to each one of the final users, in addition to integral or completes the tests that they are necessary. Besides, it must validate the integrity and security of the information stored in the DW, as well as their accessibility and availability to all its components. Also, it is required to verify the control systems and transmission of all type of information and/or data.

Activity 15: Training of the operative personnel and the final users.

This is a very complex activity and requires patience and a great capacity to transmit the concepts related to the DW, as much of operation and maintenance as of its final usage. It is necessary to remember that the number of possible presentations or options of requirements or consults can be very high and it is necessary to enable personnel who in a given time are not so close to these systems. To the operative per-

sonnel, it is required to show all the variants of extraction, filtrate, purification, integration, load or creation of the bucket of data, as well as its later operation variants.

Activity 16: Release.

Finally, after a certain use of the resources: materials and financials; as well as, great amount of work and effort, are at readiness to give the system to the area involved for its complete use. Now, we will have to monitor the performance of all the process for necessary refining or changes or common or unexpected maintenance, this will in a while lead in a specific time to a possible redesign or to even create another new similar system and thus to close the corresponding service life.

3 Conclusions

In the Area of Systems Engineering of the SEPI at ESIME, a course in the Masters in Engineering of Systems is being taught, relative to: Great design and construction of and complex systems of data bases in which it appears, in detail, each one of the activities for the construction of a DW, also includes an investigation on the existing architectures and the administrative processes of: creation, control, operation and use.

Of all these elements investigations are carried out, to know how their state present and thus to propose or to develop investigation projects or Mayor Thesis. In order to conclude these ideas a "formula" appears, that from my point of view, reflects the data concept warehousing:

Data Warehousing = change winds + Information storage + New Technologies + Complex systems.

References

1. Galindo L., "Reporte Técnico: Diseño e Implantación de Grandes y Complejos Sistemas de Bases de Datos", Maestría en Ciencias en Ingeniería de Sistemas, SEPI, ESIME Zacatenco, IPN, México D.F., MEXICO, 2004.
2. Inmon W. H., "Building the Operational Data Store", John Wiley & Sons, New York, USA, 1996.
3. Poe V., "Building a Data Warehouse for Decision Support", Prentice Hall, New Jersey, USA, 1996.
4. Galindo L., "Panorama de las Bases de Datos Federadas y Multibase", Reporte Técnico # 10, CeNaC (CIC) del IPN, México, D. F. MEXICO, 1994.
5. Bischoff J., Alexander T., "Data Warehouse: Practical Advice from the Experts", Prentice Hall, New Jersey, USA, 1997.
6. Kimball, R. "The Data Warehouse Toolkit", John Wiley & Sons, New York, USA, 1996.
7. Codd E. F. et al, "Providing OLAP (On-Line Analytical Processing) to User-Analysts, an IT Mandate", E. F. Codd & Associates, USA, 1993.