

X-META: A Methodology for Data Warehouse Design with Metadata Management

Liane Carneiro
lcarneiro@sefin.fortaleza.ce.gov.br

Angelo Brayner
brayner@unifor.br

University of Fortaleza - UNIFOR
Fortaleza - Ceará, Brazil

Abstract

Data Warehousing is a powerful tool for supporting decision-making processes in modern corporations. However, developing a Data Warehouse (DW) is a complex and costly activity. It requires strategies, which should be specific to the characteristics and needs of the organization where it will be introduced. This work presents a DW development methodology suitable to organizational environments which need Data Warehousing support as a strategic requirement, but have to use the internal staff which has no expertise in DW development. Moreover, a strategy for creating and managing metadata in an integrated way to the DW development process is proposed, in order to capture the organizational knowledge and minimize the problems which can be caused by the mobility of employees within and between organizations.

1 INTRODUCTION

A particular interest has been observed for quality information in modern corporations, even in those, which have no tradition on using computer system for decision support. On-Line Transaction Processing Systems (OLTP), though essential for performing daily operations of corporations, offers little support, if any, to decision making. Data Warehousing is becoming a solid provider of business intelligence tools that process and transform data in accurate information, fast and easily accessible. However, to promote a strategic differential in a globalize world, it is not enough to acquire technology, but also know how to use it in an effective way, joining to this the capacity for gathering, interpreting and using the information. In addition, it is important that the users of that technology can search and find information in a more logical and intuitive way. Metadata, popularly defined as data about data, aims at his functionality.

Data Warehouse Systems are Decision Support System (DSS) targeted at processing large amount of historical data, in order to identify profiles, patterns, behavior and tendencies [6] [9] [13]. Those systems have become popular. However, developing and managing such systems represent a demanding and costly activity [15], which may limit a wider adoption of this technology. Aiming at making the DW developing process more efficient, several development methodologies have being proposed. Works such as [4], [9], [13], [16], among others, claim that the DW development process should be performed by a staff with expertise on this area. It is argued in [4] that the inexperience factor is one of the main causes of failure in DW projects. However, many modern enterprises are inserted in the following scenario:

- They have no tradition on using computer systems for decision support, but managers feel necessity in that technology;
- There exists computer infrastructure for building DW and their staff is composed by professionals highly skilled in developing OLTP systems and database technology, but without expertise on developing a DW;
- The decision-makers have normally little experience about business changing, due to constant turnover in management and strategic levels.

Organizations inserted in such a scenario lack a DW development methodology that takes into an account the available staff and their experience. Moreover, in order to reduce the impact of employee

turnover and to help users to find information they need quickly, the strategy of gathering, maintaining e delivering metadata must be defined since the beginning of the project.

Considering this context, the contributions of this paper are twofold. First of all, it proposes a methodology, which addresses the development of DW projects. The proposed methodology aims at a gradual adoption of DW technology by corporations. It considers organizations with no tradition in using and developing DSS in general, and have a qualified staff, but without know-how in developing a data warehouse. Second, it presents a strategy to cope with metadata management integrated to the DW development process.

The rest of this paper is structured as follows: Section 2 analyses existing DW development methodologies; Section 3 presents and discusses the proposed methodology; Section 5 assesses the X-Meta methodology in a study case; and finally, in section 5, draws conclusions and presents future work.

2 RELATED WORK

During the last few years, much has been done with respect to the definition of DW development methodologies. Using a consistent, an efficient and a tested methodology is a key success factor of DW projects. However, the existent methodologies are normally addressed to generic environments. Thus, they do not address DW development in organizations that need a methodology adapted to their characteristics and expectations. Several works have been done in order to identify fundamental characteristics, which can guide a methodology attending all the specific needs of a corporation.

Table 1 presents a summarized comparison of the DW methodologies analyzed. On can observe that although the set of phases in such methodologies is slightly different, in fact, they are quite similar with respect to components, functionalities and activities addressed. Therefore, they have been used as a conceptual framework for the present work. The criterions “Members of the project team” and “Metadata creation and management” have to be considered as critical factors in specifying a methodology, which is able to satisfy the needs of corporations with the profile already described in the previous section.

2.1 Members of the Project Team

Several authors [4], [9], [13], [16] claim that the development process of a DW should be carried out by a staff with expertise in that technology. However, many organizations have professionals highly skilled in developing traditional systems and database technology, but with no experience in building a DW. Moreover, in government corporations, it is not always possible to hire external development teams or consultants due to factors such as legislation, cost, privacy, deadline and availability. It is important that a DW development methodology can soften these problems by giving the necessary support to the internal team for constructing a DW.

One can easily observe that the participation of users in software product design process is increasing. In such a process, the start point and the ending point is the user. At the beginning of the development, users´ needs are identified and the products are built in order to present easy interfaces to be manipulated. Besides, in DW, the built data modeling is often easy to be understood by end users. Therefore, it is important to improve and encourage the participation of end users in all the developing process. Users that never had opportunity to work with a new technology don´t know exactly what question to do. Considering users participation, as members of the project team, will make it easier gathering business metadata and help to overcome many political and organizational obstacles, with the project sponsor. Applying the strategy of “User Involvement” that brings users more closely to the software design process will allow that the DW be really built to solve specific problems of the business.

2.2 Metadata Creation and Management

In all of the analyzed approaches, the authors agree that metadata are critical for all aspects of the DW development process. Metadata should exist throughout the development process and during all the useful life of the DW. Therefore, metadata, in a DW project, play a key role.

Nevertheless, metadata management is a big challenge to many DW projects, mainly because there exists much heterogeneity among tools and products for creating and managing metadata in a Data Warehousing environment. That is because there is not in the industry a unified standard for metadata definition and interchange [11]. In spite of these obstacles, due to the importance of metadata in an analytical environment, some works have been developed to support the metadata management in DW. However, none of the works analyzed integrates metadata creation and management to the development methodology.

Table 1 – DW Methodologies

	[4]	[9]	[12]	[13]
Evolutionary approach	Yes	Yes	Yes	Yes
Members of the project team	Experienced team and business user	Experienced team	Inexperienced team	Experienced team and end-users participate in the implementation team
Product of the development process	- Data architecture - BD project - Metadata	- Data architecture - Functional architecture - Technical infrastructure - Dimensional modeling - BD project - End-User application - Metadata	- Data architecture - Functional architecture - Technical infrastructure - Dimensional modeling - BD project - End-User application	- Data architecture - Functional architecture - Technical infrastructure - Dimensional modeling - BD project - End-User application - Metadata
Detail level	Not much detailed	Very detailed	Detailed	Detailed
Pilot project	No	"proof of concept" (Product selection & installation)	"architecture and infrastructure" (Experimentation Phase)	" proof of concept " & "architecture and infrastructure", before deployment of the DW
Metadata creation and management	Not much detailed	Detailed, but not integrated to the methodology	Not detailed	Detailed, but not integrated to the methodology

3 X-META: A METHODOLOGY FOR DATAWAREHOUSE DESIGN

As already mentioned, many organizations lack a DW development methodology that takes into an account the available staff and their experience. For that reason, the key idea of the proposed methodology is to start with a DW pilot project in order to introduce experience in constructing a DW into the internal team and to prove the viability and the DW importance to an organization.

The overall methodology lifecycle is divided into 5 major phases, as depicted in figure 1. To facilitate understanding the X-Meta methodology, presents the following pattern for phase division:

Phase {Sub-phase {Group {Module {Activity
 (0,n) (0,n) (1,n) (1,n)

Thus, a phase in the X-Meta methodology may have several sub-phases, each of which may have groups. In turn, a group is composed by modules, which encompass activities.

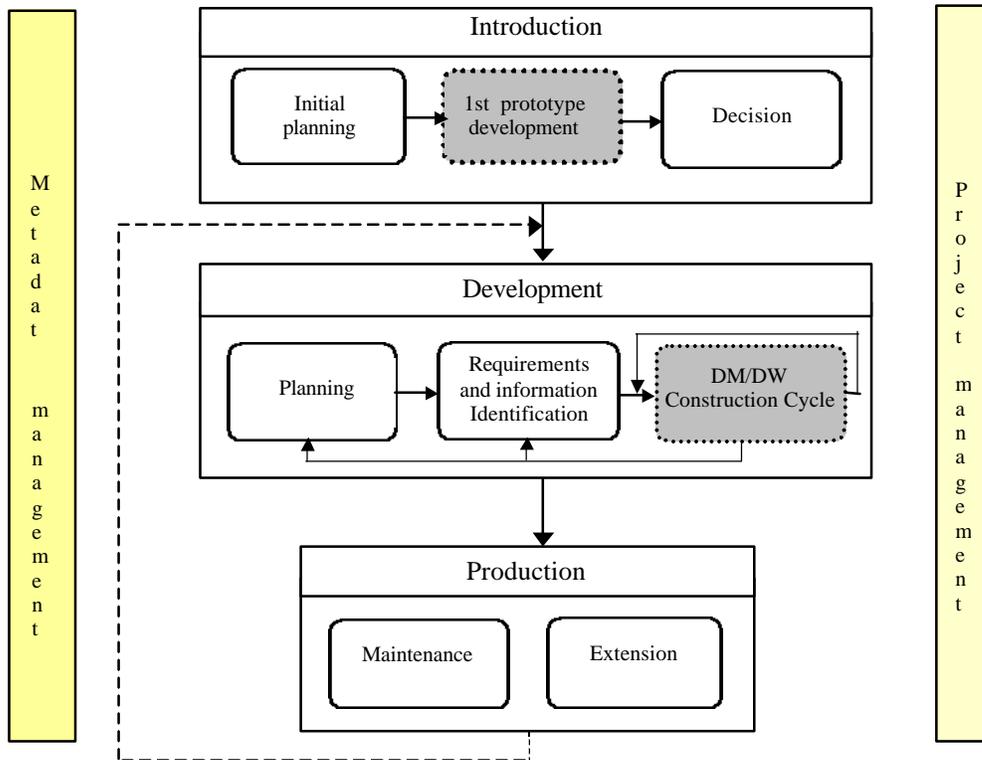


Figure 1 – X-Meta Life Cycle

3.1 The X-Meta Methodology Paradigm

The DW development process of the X-Meta methodology is based on the spiral model [14], which combines prototyping with elements of the classical software lifecycle. The goal in using the spiral model is to make easier the DW developing process, since after an iteration in the spiral model, a more complete and new DW version is generated (See Figure 2). By doing this, the user and the development staff are able to understand and react to risks in each evolutionary level. The X-Meta methodology defines three distinct iteration types, each one with its own specific purpose:

- (1) **First prototype** – it has the main goal of allowing the insertion of the DW technology in the organization. This iteration, which is executed only once in the DW development process, uses only the Introduction phase in the development of the first pilot project in the organization;
- (2) **Pilot Project** – it allows the incremental and evolutionary development of pilot projects, each of which with its own purpose, for instance: testing products, acquiring experience, developing metadata repositories, etc. Each pilot project represents an iteration, which begins in the Planning phase and uses one or more input from the DM/DW Construction Cycle phase, depending on the evolutionary level and on the project goal.
- (3) **DM/DW Project** – In this level, many pilot projects may have already been developed and many uncertainties have already been eliminated. Besides, the development process in this iteration type, which corresponds to DM/DW projects, uses the same methodology used in the pilot project iteration type, described above. Some modules in the DM/DW Construction Cycle phase may need only reviewing when the data architecture has already been defined (e.g. functional architecture). However, new peculiarities, such as greater data volume and complexity executing certain activities, will require more effort from the staff. The Production phase plays a fundamental role in this level when complete DM/DW versions are delivered.

The main idea is that the final product of the X-Meta methodology – Corporative Data Warehouse in the Organization – represents the result of the execution of many iteration types. Figure 2 presents an example of possible projects to be developed in an organization using the methodology presented in this paper. Each spiral represents a project and each project uses a specific iteration type. The first iteration (First prototype) is executed only once and the other two types (Pilot Project and DM/DW Project) can be executed many times. After the first iteration, which corresponds to the first prototype to be delivered, the next iterations will be evolutions of the first one, where prototypes can be delivered in each level. The extension of each project will depend on the purpose of the project, on the experience level acquired by the DW staff, on the time and staff availability.

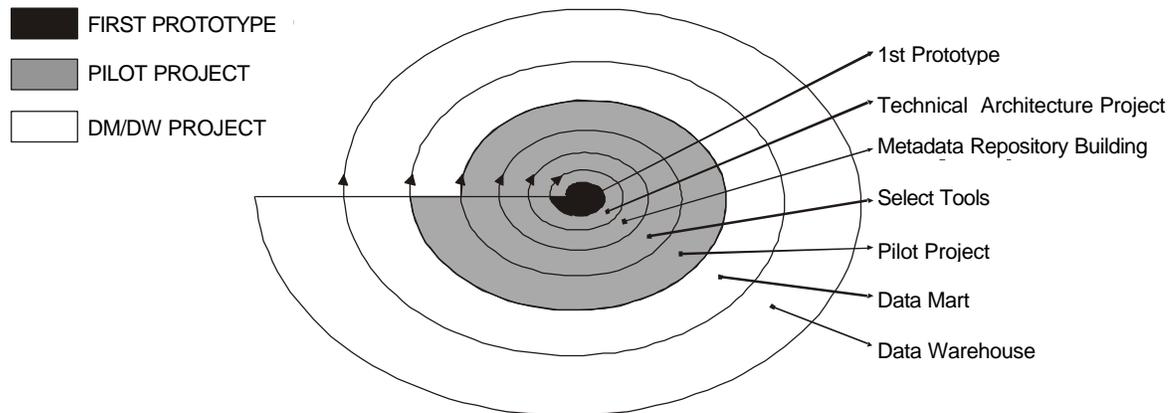


Figure 2 – Examples of iteration in the X-Meta Methodology

3.2 Introduction Phase

This phase constitutes the main differential in the proposed methodology, since there is no similar phase in DM/DW lifecycle in existing methodologies. In [13], the authors suggest the construction of a “Conception Proof” pilot project before the construction of the DW. However, such an activity is not integrated to the proposed methodology. In [12], the author presents a methodology, which addresses the construction of a pilot project. However, that activity is used to test the great variety of DW tools available.

In the X-Meta methodology, a complete method for developing a first project was defined. The idea is to allow the project to be lead by a team composed of organization staff, which has knowledge on *Data Warehousing* technology, but has not developed a DW yet. To make the first prototype available, it is necessary to have an infrastructure available in the organization.

The following benefits are expected from this phase:

- Learning the different tasks involved in the DW development;
- A first version of a Dimensional Business Model;
- Acquiring experience in activities such as mapping, extraction, transformation and loading data to the DW;
- Identifying and gathering metadata.
- Establish continuous interactivity with users, in order to gather decision requirements and provide users with useful information.

3.2.1 Initial Planning Phase

The main goal of this phase is to guarantee that, even for a project with reduced scope, the main activities of a system development project are executed efficiently. The collected information and the obtained results will be useful to the next phase and to plan future projects. This phase is executed in six independent activities, shown in table 2.

Table 2 – Initial Planning Activities

Activity	Goals
Define the project scope	Give priority to a business area with great impact in the organization and establish interactivity with users.
Define the technical infrastructure	Specify the technological resources (hardware, DBMS, tool for presenting data) available in the organization to be used in the project.
Define the participants, responsibilities and deadlines	Determine the project deadline and create a project plan.
Define the selected business area requirements	Understand and collect the selected business area requirements.
Define the information made available to users	By identifying the types of reports used by the decision makers, define the reports and priority queries that will be made available.
Define the metadata solution to the prototype	Identify, create and use the metadata in the prototype.

3.2.2 1st Prototype Development Phase

To develop a first version of the DW prototype, only the third entering point of the phase DM/DW Development cycle (represented by ③ in figure 4) has to be used, as it can be observed in figure 3.

Although the modules in this phase have similar names and activities to the modules in the third input group in the DM/DW Construction Cycle phase, which is part of the Development phase, illustrated in figure 4, the activities in this phase are simpler in order to make the prototype available faster, not interfering in the reliability and quality of the produced DW.

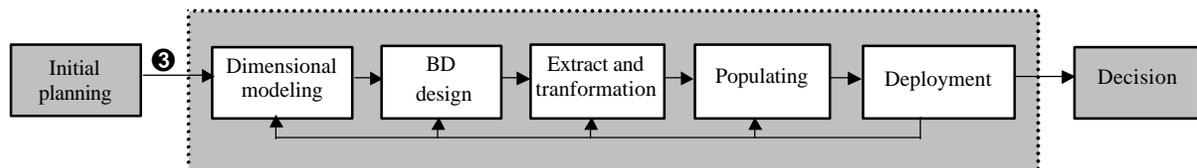


Figure 3 – Development of the 1st prototype

3.2.3 Decision Phase

After delivering the prototype, comes the most uncertain moment in the project because of the decision expectation – Invest in a corporative DW or forget about this idea, which is a decision that will be made by the organization managers at the end of this phase. The main activities are the support to users when using the first prototype, communication and publicity of the success obtained by using the DW.

With the final project acceptance, it is considered that the DW technology was successfully introduced in the organization and that a new challenge will take place, developing a more complex DW.

3.3 Development Phase

The Development phase aims at the effective development of DW projects, mainly of Data Marts (DM), referred to as a subject-oriented DW. Even in pilot project all the phases of this phase must be used, enabling to acquire experience and to eliminate or reduce, as much as possible, most risks and uncertainties that normally jeopardize the development process; also to transform the acquired experiences in more precise definitions.

The specifications of the phase “Planning” and “Business requirement definition” are adaptations of [9].

3.3.1 DM/DW Construction Phase

The DM/DW construction phase (figure 4) is the main phase of the Development phase and it constitutes a set of development modules. These developments modules are organized into groups, depicted in figure 4 by ❶ ❷ ❸ ❹. Such groups represent different entering points to this phase. The existence of these different points allows that a specific project can be developed using only a group or two groups, or all of the groups.

The requirements and definitions originated from the phases describe before (“Planning” and “Requirements and information identification”) are used as input to this phase and can be used by any group. These groups can be executed concurrently or individually. The last modules of those groups is the Deployment module. The results delivered to the Production phase may not happen in certain pilot projects or even have a more restrict use. For instance, in a pilot project with the goal of testing and selecting an OLAP tool, the resulting product of this project may not be directly made available to the end user, but it may be incorporated to the project of technical architecture definition of the DW environment in the organization. It is not necessary to execute the Production phase in this case.

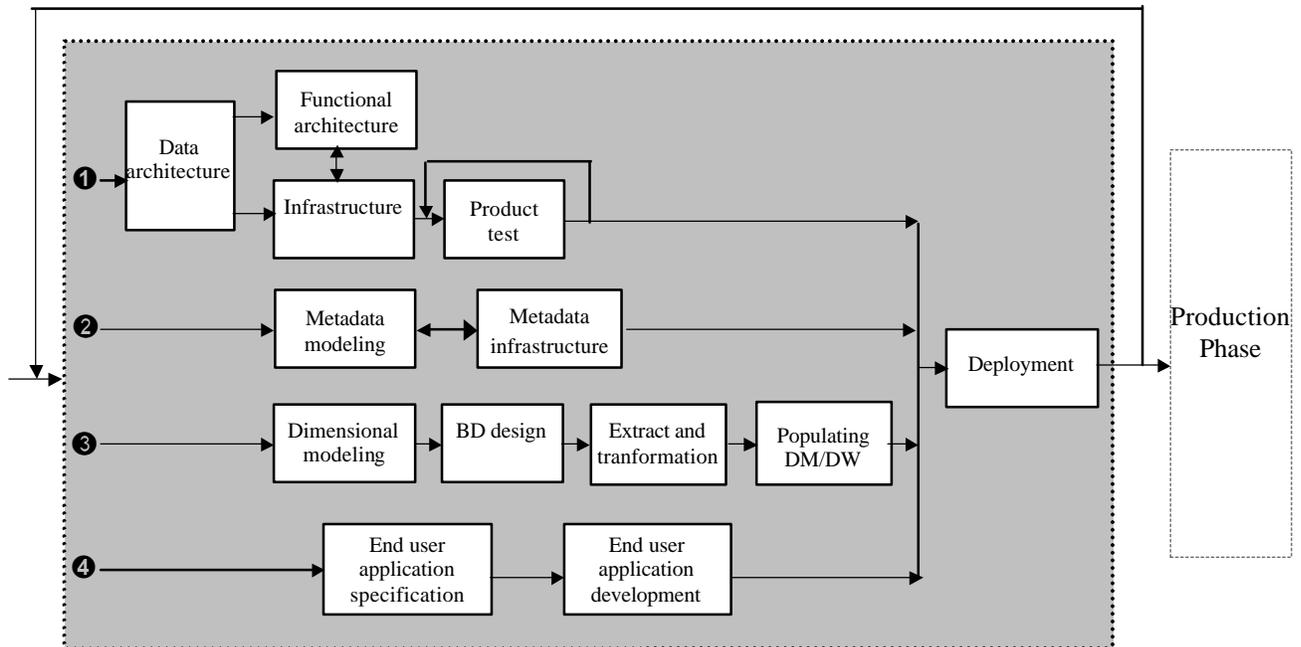


Figure 4 – DM/DW Development Cycle

The goals and activities of most of the modules that constitute this phase are adaptations of other works, especially [9], [13]. The modules that constitute the second input in this phase (❷ in figure 4) address the metadata management as an integrating part of the DW development and production process. Next, the goals and activities of this group are briefly discussed.

Metadata modeling

It has the goal of defining and controlling the activities related to the metadata, supplying corporative directions for its creation, use and maintenance. A complete solution to the metadata management, including the metadata for developing and controlling the production environment changes may be difficult in the beginning and it is time and money consuming. In [13], the authors recommend managers to start with simple solutions and evolve into metadata solutions for data integration, which is more difficult to build for the first DM/DW.

Following, there is a list of activities to be performed in this module. Those activities have the goal of producing a metadata model that will be used during the construction of a metadata repository and during the metadata management phase:

- Define the metadata management team and its responsibilities;
- Define and classify metadata types that will be stored;
- Define user types a repository and the equivalent access levels;
- Determine the metadata sources in the organization
- Define and construct the metadata model and metadata flow.

Metadata infrastructure

The metadata repository is an infrastructure component of the DW environment that is useful to all other environment components, working as a tool to help in the integration. The goal of this module is to define the general metadata infrastructure in the organization from the definitions in the previous module (“Metadata modeling”).

The activities belonging to this module are:

- Construction of an architecture for the metadata (centralized, decentralized, distributed);
- Evaluation of metadata tools in order to select the best one to achieve the project goals;
- Physical implementation of the metadata repository;
- Defining security procedures.

The repository development and a strategy to collect, maintain and distribute the metadata should be defined. In addition, make sure that a mechanism that populates and maintains the metadata repository and that the access to the DW has metadata as input.

3.4 Production Phase

The specifications of this phase include the necessary tasks to support the ongoing growth of the DW. Its specifications are adaptations of [9] [13] [16]. A strategy for shared metadata from the latest phase and this phase will reduce maintenance costs and update cycle.

4 CASE STUDY

SEFIN is a department of the Fortaleza City Hall, which has the goal of “Developing the financial, budget, tax and fiscal policies, along with the Municipal Executive Power”. Fortaleza is a city located in Brazilian northeast in which lives more than two millions people. The collection business area is responsible for the incoming of financial resources in the Municipal safe, which are the income collected from the IPTU (Urban territorial and land Tax) and ISS (Tax over services) taxes, among others. SEFIN has highly qualified professionals; some of them are familiar with decision support technology, although they lack practical experience. Some administrators do not have enough knowledge on this business, since changes in the business rules and in the direction position are very common in Brazil.

The SEFIN OLTP systems are composed of 230 files and 800 Dataflex applications developed in 1990. Administrators and decision-makers feel an enormous need of managerial and tactical information, especially information on the changes in the business rules over the years.

Before starting the work, an analysis based on the viability of the project's accomplishment, approaching costs and benefits, was presented to the Financial Secretary, who approved the project and assigned a staff for it. Three technicians composed the internal development team. The team was composed by a manager and a developer, both involved with the Sefin OLTP system and having theoretical knowledge on DW technology, but no practical experience on the development of such systems; and one database administrator, having no knowledge on DSS. The user who analyzed the results of the IPTU collection joined to the staff.

The application of the X-Meta methodology in this project was fulfilled through the application of the Introduction phase, when all the activities described in section 3.2 were executed.

Through interviews and facilitation sessions, the staff defined which reports, proceeding from the OLTP Systems, were manipulated, defined the analytical needs, and defined the reports and priority queries to be made available by the prototype. After the analysis of all the identified information, the business dimensional model was projected, focusing on the data granularity. The DW logical project and its physical implementation in the Oracle 8i Database were activities developed along with a DBA, when the prototype functional architecture was completely defined. The DW availability towards the business analysts happened through the Excel spreadsheet, which accesses the DW and allows users to create different views for the information available. The metadata solution in the prototype was also finished and users could consult the information available manually.

The following aspects were observed during the execution of the activities performed during the development of the 1st prototype:

- The technical resources made available (hardware, software, etc) did not have any financial investment;
- Available reports, usual decision procedures, involved OLTP systems (including the Dataflex code), etc, were studied;
- Users expectations in executing the DW project grew as each activity was executed and presented.

It was proved that the main goal defined in the beginning of the project was accomplished, which is to introduce the *Data Warehousing* technology at SEFIN. It was possible for an inexperienced staff to construct a DW prototype and generate correct information in the right moment to the decision makers, by executing only the first phase of the methodology presented here. In addition, this staff was able to obtain clear information about the changes in business rules that happened over the last five years, by using metadata, allowing a complete analysis over the information collected. The results obtained by using the prototype were better than expected, shown through the easy and fast way to obtain managerial information, what surprised the administrators tremendously.

5 CONCLUSIONS

In this paper, a DW development methodology was described and analyzed. The proposed methodology, called X-Meta, addresses the problem of developing a first DW project in corporations, which do not have staff with practical experience in such development. Furthermore, the X-Meta methodology integrates metadata creation and management to the DW developing process.

The proposed methodology is being used to construct a DW for a department of the Fortaleza City Hall. The first prototype has been already developed, without any financial investment and using resources available in the organization.

As future work, we envision the following activities:

- Validating the methodology in other contexts of use;
- Applying the methodology in pilot projects for complete DW systems;

REFERENCES

- [1] Bontempo, C., Zagelow, G. **The IBM – Data Warehouse Architecture**. Communications of the ACM, v. 41, n. 9, p. 38-48. September, 1998.
- [2] Chaudhuri, S., Dayal, U. **An Overview of Data Warehousing and OLAP Technology**. ACM SIGMOD Record v. 26, n. 1, p.65-74, March 1997.
- [3] Data Warehousing Tools Bulletin. **What is Metadata**. ComputerWire, 1996. URL: http://www.computerwire.com/bulletinsuk/212e_1a6.htm. In December, 2000.
- [4] Gardner, S. R. **Building the Data Warehouse**. Communications of the ACM, v. 41, n. 9, p. 52-60. September, 1998.
- [5] Gray, P., Watson, H. J. **Decision Support in the Data Warehouse**. New Jersey, Prentice Hall PTR, 1998.
- [6] Inmon, W. H. **Building the Data Warehouse**. New York:John Wiley & Sons, 1996.
- [7] Inmon, W. H.; Welch, J. D.; Glassey, Katherine L. **Managing the Data Warehouse**. New York:John Wiley & Sons, 1998.
- [8] Keen, P. G. W., Morton, M. S. **Decision Support System: an organizational perspective**. Addison-Wesley Publishing Company, 1978.
- [9] Kimball, R., Reeves, L., Ross, M., Thornthwaite, W. **The Data Warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouse**. New York:John Wiley & Sons, 1998.
- [10] Kimball, R. **The Data Warehouse Toolkit**. New York:John Wiley & Sons, 1996.
- [11] Marco, D. **Building and Managing the Meta data Repository – A Full Lifecycle Guide**. New York, John Wiley & Sons, 2000.
- [12] Pereira, W. A. L. **A Methodology Targeted at the Insertion of Data Warehouse Technology in Corporations**. MSc. Dissertation. Porto Alegre-PUCRS, 2000 (in Portuguese).
- [13] Poe, V. Klauer, P.; Brobst, S. **Building a data warehouse for decision support**. New Jersey, Prentice Hall PTR, 1998.
- [14] Pressman, R. S. **Software Engineering: A Practitioner’s Approach**. McGraw Hill, 1997.
- [15] Sen, A., Jacob, V. S. **Industril – Strenght Data Warehousing**. Communications of the ACM, v. 41, n. 9, p. 29-31. September, 1998
- [16] Watson, H. J. **Managerial Considerations**. Communications of the ACM, v. 41, n. 9, p. 32-37. September, 1998.