A DATA WAREHOUSE ARCHITECTURE FOR BRAZILIAN SCIENCE AND TECHNOLOGY ENVIRONMENT

André Luís Andrade Menolli, Maria Madalena Dias
Department of Computer Science, University of St. Maringá ,579 Colombo Av., 87020-900, Maringá, Brazil

Keywords: Data Warehouse, Data Integration, Science & Techology

Abstract: Science and technology in Brazil are areas that have few available resources and many times these scarce resources are badly used. The data warehouse is a tool that can make possible a better distribution of these resources. In this article are considered some issues in the development of a data warehouse for Science & Technology management. The paper describes the necessity of a supporting system to the decision taking regarding the distribution of the resources destined to Science & Technology in Brazil, and also shows a data warehouse architecture that is being developed to support this system. data modeling characteristics defined for the proposed data warehouse architecture are presented too.

1 INTRODUCTION

The data warehouse (DW) technology has been widely used in companies with the aim of offering organization, management and database integration. In the data knowledge discovering process, the first step is the data preparation, where the data must be organized and stored in the DW.

The research in DW area had a big increase in the nineties, mainly after 1996. This can be verified in the study made by Vassiliadis (2000).

Despite this increase in DW area research, it is still very difficult to have support in developing a DW , because the research are still turned to the academic area. The gap between a practical DW and a research one became obvious (Vassiliadis, 2000).

Another important point is the lack of project and structure in the DW development. According to Demarest (1997), the DW project has some failure factors, such as:

- Problematic data engineering.
- Unrealistic schema design.
- No design method is used.

Due to these reasons it is of extreme importance to have an architecture project with well defined steps assisting the DW designer in the conclusion of the project successfully and in the expected time.

The main objectives of the work presented in this paper are the definition of an architecture and the construction of a DW for the Brazilian Science & Technology system.

The construction of this DW brings as benefit a better acquaintance of the brazilian research reality, making easy the investments exploitation in this area.

Therefore, the DW is used as a knowledge extracting tool, with the objective of reducing cost and optimize the quality of its products and services (Dias et al., 1998), thus obtaining more and better results with less resources.

The remainder of this paper is organized as it follows: Section 2 presents a brief vision of Science & Technology in Brazil, as a motivation for the development of this work. Section 3 presents the developing proposed DW architecture for the project. Section 4 presents the data modeling that was used in the staging area and in the DW. Section 5 concludes this paper.

2 SCIENCE AND TECHNOLOGY IN BRAZIL

In Brazil or any other developing country, a key-point is to optmize scarce resources destined to Science & Technology, minimizing wastefulnesses and making the expenses follow a previous planning. An alternative to obtain this, according to Romão (2002), is to increase the investments in Research and in Development, getting to know in details the infrastructure and potential reality of

Luís Andrade Menolli A. and Madalena Dias M. (2004).
A DATA WAREHOUSE ARCHITECTURE FOR BRAZILIAN SCIENCE AND TECHNOLOGY ENVIRONMENT.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 522-525
DOI: 10.5220/0002604305220525

research in the country, as well as researchers and research profile. To obtain more and better results from little resources is a great challenge for developing nations as Brazil.

The brazilian Science & Technology environment possesses important databases that are consolidated and trustworthy, from which is possible

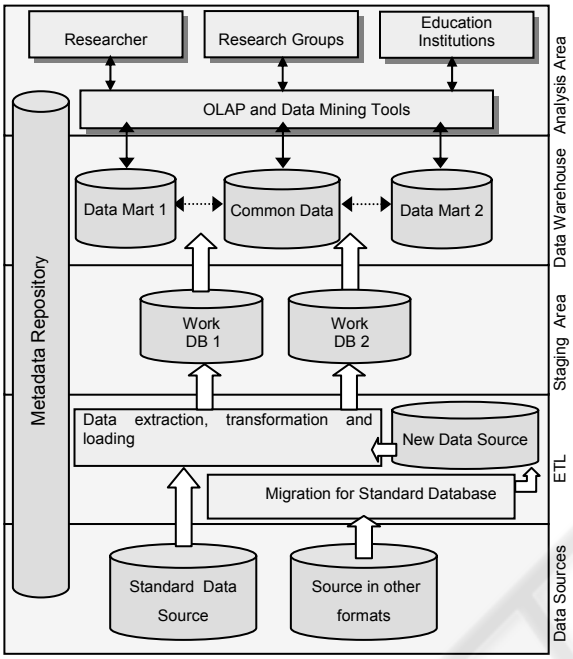

Figure 1: The project DW architecture

to trace a Science & Technology panorama in any region of the country.

The Science & Technology evaluation is a process tied to the promotion, that is to say, it is necessary an analysis that produces subsidies for the decision taking, determining the levels of resources distribution. To carry through any type of information analysis, it is necessary to use complete information that represents on a concise form the reality that is intended to analyze.

That being so, the DW projected has as its objective to serve as a support for knowledge discovery on Science & Technology that can be used to analyse the researchers, research groups, courses and institutions performance, by identifying which are the areas that are bringing investments back and which need new investments.

3 THE PROPOSED DATA WAREHOUSE ARCHITECTURE

The architecture strategy of this project is mainly to replace data sources of diverse formats for new standardized analytical databases, facilitating then, the data integration. The architecture is constituted of several layers, as it's shown in Figure 1. The thick arrows represent the data shipment while the fine arrows represent the data access. The architecture is constituted of five layers, so that the layer 2 can be subdivided into two other layers, depending on the data source type existing in layer 1.

3.1 Data Sources

This layer stores the original data sources, which could be in two formats:

- Standard Format.
- Distinct Formats.

A standard database was chosen, that in this project is a relational normalized data model and data in this format are in the standard format. Data that are in another type of database, or in other formats (files, text, etc), are migrated to the normalized model in the chosen standard format.

3.2 Extraction, Transformation And Loading (ETL)

The first step in this layer is the verification of the data source format. If the data are not in a standard format, the data source will go through a process of data migration, where the data source will be integrated and transformed into the standard format. This process will facilitate the data manipulation, data consistency and data integration.

3.3 Staging Area

According to Dumoulin (2003), a fundamental concept that greatly simplifies DW projects and ongoing maintenance is the use of data staging areas.

From the staging area logical project, data modellers have a good idea of the necessity of attributes and sources to populate the DW. The data staging area serves as a dividing line between the systems sources and the DW, containing only information necessary to populate warehouse. The staging area defined in this work uses a normalized data model, to allow a high data consistency, and to facilitate the integration process among distinct databases.

3.4 Data Warehouse

The DW is the storage area and follows the model approach proposed by Kimbal et al (1998), using the BUS implementation, which needs standardized dimensions and facts. This project uses de

dimensional modeling, because according to Song et al (2001) there are two main advantages in the use of dimensional models in DW environments. Firstly, a dimensional model provides a multidimensional analysis space in relational database environments; we are analyzing factual data using dimensions. Secondly, a typical denormalized dimensional model has a simple schema structure, which simplifies the final user query processing and improves performance.

3.5 Analysis Area

In this layer, consult and mining tools can be utilized, although they are not in the scope of the present project. These tools will access the data marts or materialized views, based in metadata repository.

4 MODELING PROJECT DATA

The modeling process used in our project is based on the classical database modeling process that distinguishes the conceptual, logical and physical models extended by the implementation model approach proposed by Kimball et al (1998).

From the data sources until the data in the formats that appear in the model, showed in the Figure3, the following steps are executed:

1. Separating the data by granularity.
2. Cleaning values of attributes.
3. Transforming the data.
4. Changing the operational keys by substitute keys.
5. Guaranteeing the quality of the data.

The use of these steps give the follows benefits to the data: quality, standardization and consistency. Thus facilitating the data integration and the data load in the DW.

The DW was projected according to the star schema. This modeling has been chosen for the already presented advantages in this paper and because dimensional modeling is widely accepted as the viable technique for delivering data to end users in a DW (Kimball et al, 1998), (Meyere & Cannon, 1998), (Axel & Song, 1997), (Dinter et al., 1997).

However, the star schema does not accept many-to-many relationships, and this type of relationship presented in the staging area of our project, will be transformed into many-to-many relationships between facts and dimensions in the DW. According to Song et al (2001), those relationships in a dimensional model causes several difficult issues, such as losing the star schema structure, increasing complexity in forming queries, and degrading query performance by adding more joins. Therefore, it is desirable that we handle the many-to-many relationships while still keeping the structure of the star schema. There are some works to resolve this problem, as (Krippendorf & Song, 1997), (Theodoratos & Sellis, 1998) (Lehner, Albrecht, & Wedekind, 1998), (Pedersen & Jensen, 1999), (Kimball et al, 1998). Among these solutions, the chosen one was presented by Kimball, because it is ideal for dimensions with no upper limit in many side and it is a clean solution (Song et al., 2001), that is adjusted to this project.

These many-to-many relationships between facts and dimensions and the relationships between dimensions and mini-dimensions presented previously, can be seen in Figure 3. In this figure is also possible to verify that more than one fact uses

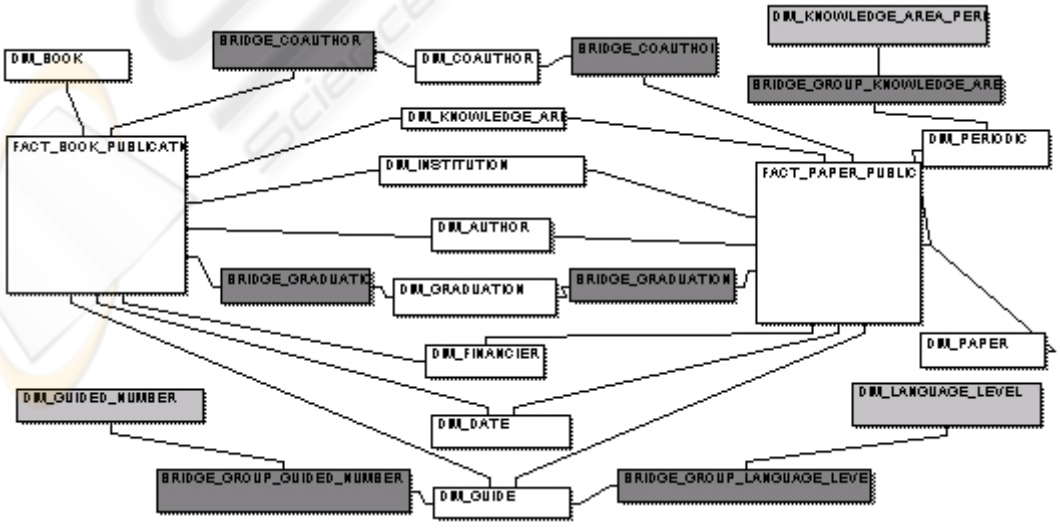

Figure 3: Partial modeling of DW.

the same dimension, according to what was proposed with the use of the DW Bus architecture matrix.

The use of the star schema altogether with many-to-many techniques of modeling between facts and dimensions and among multivalued dimensions, it is expected to obtain a simplified and denormalized dimensional model, improving the inquiry processing of final users and raising the performance.

5 CONCLUSIONS

In this project it was defined a DW architecture giving support to the decision taking in Science & Technology in Brazil. This architecture is divided into diverse independent layers, facilitating the resolution of problems in each one of the DW development stages. In this architecture, diverse DW technologies and classical relational database technologies are used. The DW data modeling was defined considering that the knowledge to be obtained is related, mainly, to the Brazilian researchers scientific production .

It is expected that with the proposed DW in the project, it is possible to minimize existing problems in the resources distribution destined to the Science & Technology in Brazil.

REFERENCES

Axel, M., & Song, Y. (1997). Data Warehouse Design for Pharmaceutical Drug Discovery Research. *Proceedings of 8th International Conference and Workshop on Database and Expert Systems Applications (DEXA97),* 644-650, Toulouse, France.

Dias, M. M., Mattos, M. M., Romão, W., Todesco, J. L., & Pacheco, R. C. S. (1998). Data Warehouse -Presente e Futuro. *Proceedings of Revista Tecnológica*, 7, 59-73, Brazil.

Demarest, M. (1997). The politics of data warehousing. Retrieved September 10, 2003, from http://www.dmreview.com/whitepaper/wid293.pdf

Dinter, B., Sapia, C., Hofling, G., & Blaschka, M. (1998). The OLAP Market: State of the Art and Research Issues. *Proceedings of Int'lWorkshop on Data Warehousing and OLAP*, 22-27, Washington, USA.

Dumoulin, R. (2003). Architecting Data Warehouses for Flexibility, Maintainability, and Performance. Retrieved August 9, 2003, from http://www.olap.it/Articles.htm

Kimball, R., Reeves, L., Ross, M., & Thornthwaite, W. (1998). *The Data Warehouse Lifecycle Toolkit.* New York: John Wiley & Sons, Inc.

Krippendorf, M., & Song, Y. (1997). Translation of Star Schema into Entity-Relationship Diagrams. *Proceedings of 8th International Conference and Workshop on Database and Expert Systems Applications (DEXA97)*, 390-395, Toulouse, France.

Lehner, W., Albrecht, J., & Wedekind, H. (1998). Normal Forms for Multidimensional Databases. *Proceedings of SSDBM, 63-72.*

Meyere, D., & Cannon, C. (1998). *Building a Better Data Warehouse.* Prentice-Hall.

Pedersen, T. B., & Jensen, C. S. (1999). Multidimensional Data Modeling for Complex Data. *Proceedings of 15th ICDE*, 336-345, Sidney, Australia.

Romão, W. (2002). *Descoberta de conhecimento interessante em banco de dados sobre ciência e tecnologia.* Doctoring Thesis. Retrieved May 24, 2003, from Santa Catarina Federal University, Production Engineering Postgraduation Program Web site: http://teses.eps.ufsc.br/defesa/pdf/3079.pdf

Song, Y., Rowen, W., Medsker, & C., Ewen, E. (2001). An Analysis of Many-to-Many Relatioships Between Fact and Dimension Tabels in Dimensional Modeling. *Proceedings of Int'lWorkshop on Design and Management of Data Warehouse(DMDW2001),* Interlaken, Switzerland.

Theodoratos, D., & Sellis, T. (1998). Data Warehouse Schema and Instance Design. *Proceedings of 17th International Conf. On Conceptual Modeling (ER98),* 363-376, Singapore.

Tombros, D., & Häberli, C. (2001). A Data Warehouse Architecture for MeteoSwiss: An Experience Report. *Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW2001)*, Interlaken, Switzerland.

Vassiliadis, P. (2000). Guliver in the land of data warehousing: pratical experiences and observations of a researcher. *Proceedings of the International Workshop on Design and Management of Data Warehouses (DMDW'2000)*, Stockholm, Sweden.