A Design and implementation of a data warehouse for research administration universities

André Flory[1], Pierre Soupirot[2], and Anne Tchounikine[3]

[1] CRI : Centre de Ressources Informatiques INSA de Lyon - Bâtiment Léonard de Vinci
[2] LISI : Laboratoire d'Ingénierie des Systèmes d'Information INSA de Lyon - Bâtiment Blaise Pascal

Preface

Modern businesses use a great number of different computer systems to manage their daily business processes. These systems are knew as "operational system" and have been acquired from several vendors during a long time and are based on different technologies and softwares. The integration of there different systems is complex and difficult. The data warehousing approach solves the problem by integrating data from the operational systems into one common data store : the data warehouse.

The primary concept of data warehousing is that the data stored for business analysis can most efficiently be accessed by separating the data from the operational systems.

The reasons to separate the operational data from analysis data have not significantly changed with the evolution of the data warehousing systems. Advances in technologies and changes in the nature of business have made many of the business analysis processes much more complex and sophisticated. In addition to producing standard reports data warehousing systems support very sophisticated online analysis including multi dimensional analysis. [INM96CHA97]

The paper is outlined as follows section 2 describes the French university system and research structures. The section 3 shows how research administration could benefit of data warehousing solutions. The section 4 describes the architecture of a data warehouse for research administration. Section 5 summarises the paper and offers suggestions for next steps.

1 The French academic system

We can find two kinds of education institutions in France : there are about 880 "classical" universities (their type is the same that other universities in the world) and about 80 specific institutions named "high school" where the number of students is limited. That mainly concern scientific and engineering schools.

INSA de Lyon (National Institute of Applied Sciences) is an institute of technology on a European scale. It has 4 000 students and more than 800 graduates per year in ten different specialised fields (computer sciences, telecommunications, biology, . . .).

There are 30 research laboratories which works with the French major research organism as the National Institute for Agronomic Research, the National Institute for Health and

Medical Research and the National Centre for Scientific Research. These laboratories receive PhD students from the 24 nationalities accredited PhD programs of INSA de Lyon.

As all education institutions, the INSA de Lyon has an administrative information system and an academic information system which are not associated. This makes unused administrative researches available for education and research and not facilitates cross-fertilisation.

Administrative information system mainly concerns student management, employee management, payroll, … but does not concern research management : this comes from the fact that research information is present in many different databases but is not centralised in one specific application.

Our purpose is to build a system to evaluate research activity of laboratories. This application is important for :

To know the activities of laboratories ;

To have a better visibility of laboratories management ;

To detect changes : old research subjects and new projects ;

To evaluate research it is necessary to use a number of informations which are present in various database like :

Staff (Professors, Assistant Professors) ;

Research contracts with industry ;

Thesis (subjects) ;

Publications (authors, subjects, importance and place of publications).

We must select pertinent informations and extract them in several heterogeneous systems. We must also analyse all dependencies.

2 Benefits of a data warehouse solution

2.1 Definition

Following Bill Inmon [INM96] a data warehouse is "a subject oriented, integrated, non volatile and time-variant collection of data in support of management's decisions". The main objective is to bring together the various data which are distributed in transactional heterogeneous databases into a single collection [WID95], [KIM96], [CHA97]. This centralisation of voluminous amount of data allows the discovering of trends and provide hidden information that ease managers' decisions.

Figure 1 shows the main principles of a classical data warehouse (DW) architecture:

The alimentation stage (1) consists in

extracting data from transactional databases,

transforming data (filtering, cleaning, mapping and ordering)

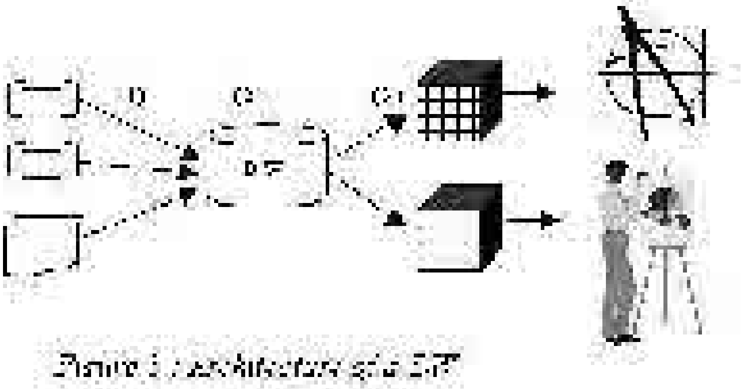


Figure 1 : architecture of a DW

loading data into the DW.

Data in the DW (2) are conceptually modelled following a multidimensional model (star or snow-flake model) and logically modelled following a ROLAP or MOLAP model [VAS99]. Multidimensional models allows to highlight facts (measures) which represent useful indicators for the managers, and dimensions which represent interrogation axis. For example, a measure for a sale activity could be the total amount of purchases (facts) calculated in relation with a type of products, geographic situation, and/or period of the year (dimensions). Physically, data are indexed, and can be aggregated, fragmented in order to optimise exploitation processes.

OLAP (On Line Analytical Processing) tools (3) then are used to process operations on the data stored in the DW. These tools operate on the multidimensional view of the data (hyper cube) and provide roll up, drill down, slice… operators allowing the end-user to navigate along and inside the hyper cube. Other exploitation operations can be applied on data warehouse such as data mining techniques implementing knowledge discovery.

Data warehousing is today very much used in business field. Many industrial products are successfully marketed : ETL (Extraction-Transformation-Loading) tools, dedicated DBMS and OLAP systems [OLA]. In Figure 2, one can see that in France, the public sector is backward business fields in data warehousing.

We try to show in the following how data warehousing can supply solutions to the goals described section 2.

2.2 Facts and dimensions

One of the first challenges in the data warehousing process is to select the facts and dimensions that are significant for the end-users. In the context of an investigation on research activity, it is difficult to determine unbiased indicators. However, one can consider that the number of scientific publications, the number of projects or contracts can describe part of a researcher 's activity and thus could be modeled as facts in a DW. Subjects such

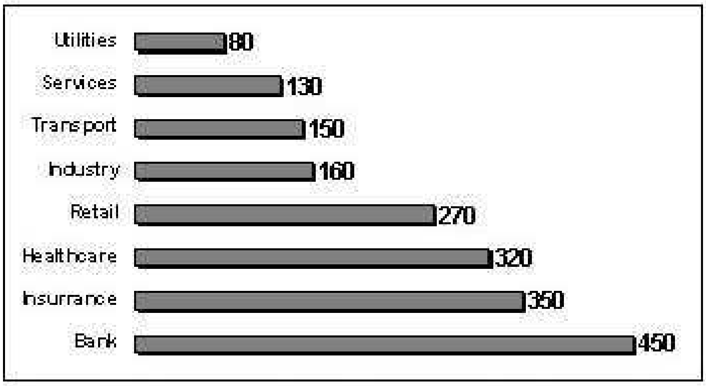


Figure 2 : Average size of data warehousing projects (thousand dollars) in France (ICD 97)

as topics, types of publications (books, article in proceedings, in journals...) or type of projects (national, European...) are relevant dimensions. We will have to deal with the fact that an article for example can have several authors, from the same lab, or not. This article should count for one article at the lab level, but for each author at the individual level.

2.3 Historization

Other interesting measurements are number of PhD students, number of associate professors and professors. The analysis of the researchers' professional career is also a good indication on the activity of a laboratory : how is the flow of persons entering or leaving a lab ? How long lasts a thesis ? Do the students become permanent researchers after their PhD degree ? What are the prospects for the members of a lab ? Conventional operational databases offer a instantaneous snapshot of data ; in a DW, data are non-volatile, and are stored for long periods. This historization allows to investigate on data evolutions along time and will be a great support to follow the career of members inside and outside of their lab.

2.4 Data centralisation and integration

Useful data for the DW are distributed in various and heterogeneous databases. Part of these data are stored in administrative databases such as in the employees or students offices. Another part lies in the laboratories. Indeed, in each laboratory, several software tools are used for daily management : staff lists or bases, that can contain the same information as in the employee office (name, surname, ...), but also complementary information (e-mail, ...), and omitted information (private address, social security information...); Lab frequently use bibliography tools to handle their publications. These tools can be marketed ones, or be delivered by university or research center, or even be developed by someone in the lab. Lastly, financial tools contain data on the contracts and projects (industrial contracts, supported projects....). ETL tools for data warehousing provide attractive solutions for centralizing and integrating distributed and heterogeneous data.

3 The design of the data warehouse

3.1 Implementation of the data warehouse

The activity of a laboratory can be measured by two parameters : the first one is the number and the quality of publications and the second one is the amount of money the laboratory receives for research contracts.

This leads us to build our data warehouse in two parts : the first one concern publications and the second accountancy.

3.2 Conceptual schema for the publication domain

The central entity of this schema is the production entities, which represents the fact table. A production is an association between a researcher and a paper (a researcher is associated with many papers and a paper may have many authors).

In the production table we only find a number which specifies the order of the researcher's name in the list of authors.

The production table represents an association between a paper and a researcher. This table describes the characteristics of the publication. In this table, we can find research papers, contract reports or PhD thesis. We represent in this table attributes as paper number, researcher number and key words.

The book table describes the material representation where the article is published. It can be a book or acts of a given congress (for example EUNIS 2001) or a publisher (Springer, Prentis hall, . . .). The book table contains the following attributes :

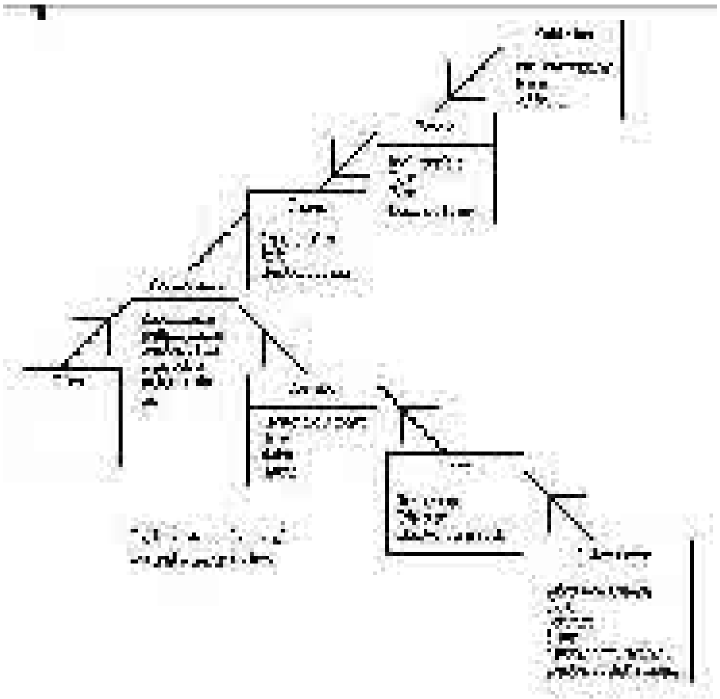
book number, type, year, number of pages.

A publisher represents a book company (ex : Prentis Hall) or an institute which organises conferences (ex : The Data Warehouse Institute). The publisher table has three attributes : publisher number, publisher name and publisher address (or town).

A production is also associated with one and only one author who is a researcher : so the production table is associated with the researcher table. We have many characteristics for representing researcher as name, status, salary, age, . . .

A team is a part of laboratory, which works on given subjects. A researcher manages a team. The team table has attributes as team name, director, number or researchers.

The top level of this dimension is represented by the laboratory : information about laboratory are represented in this table (name, director, status of laboratory, number of researchers, number of PhD students, . . .).



3.3 Conceptual schema for the accounting domain

The second part of the data warehouse deals with the accounting system and its representation. One of the main activities of laboratories is to work on public or private contracts.

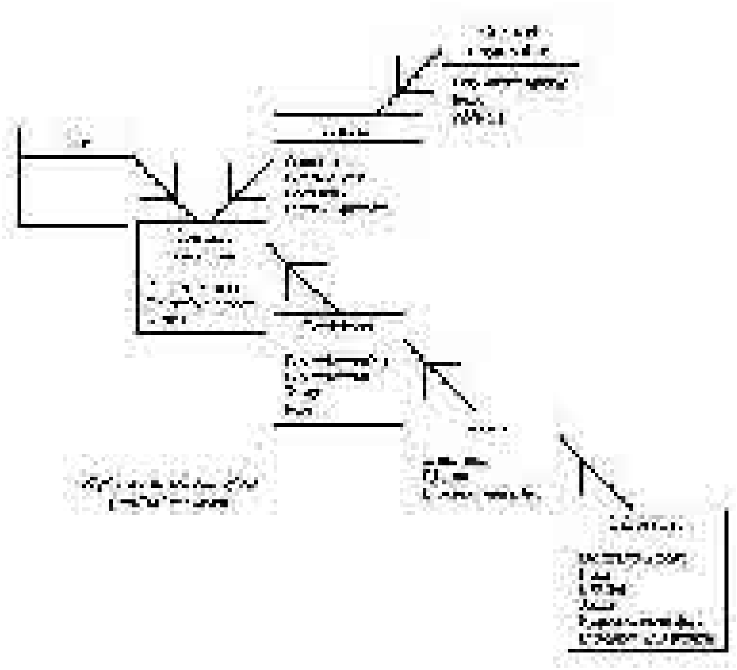
It is important to know which researcher works on a given contract and what part of his time he gives to this contract (in fact we represent a percentage of this time) : this can be half time for academic staff and full time for non academic staff but also a researcher can work in a same time on more than one project.

We see in fig 4 a star schema of the accounting system. The fact table is named "contract researcher".

Contract organisation represents a company or a government or a state ... The contents of the fact table we "measured" values of the research business.

The table is associated with the contact organisation table, which contain information about companies or government (states) offices, which pays for contracts. In some special cases, the contract organisation can be the university or the laboratory itself. So a contract is always associated with a contact organisation. The contract organisation has several attributes like organisation number, name, address, ...

The other dimensions of fig4 is the same that the corresponding dimension of fig3 where participant, as author, is another role played by a researcher.



3.4 The integration

The two systems described above represent the facts, which materialise the activity of a laboratory : publications and research contracts. The remainder problem is that there is no connection between the two systems. We only can relate the two systems through the papers. We can consider that a paper is written during the research done for a given contract.

We can link paper and contract in the following way and we consider that a paper is always written according to the work done on a given contract.

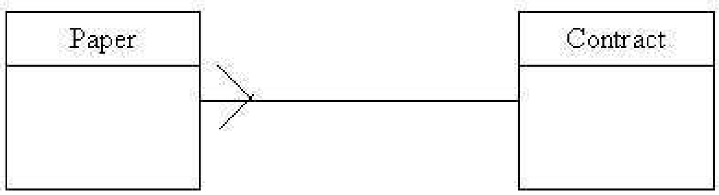


Fig 5

3.5 Data extraction

We use specialised data extraction tools to extract data from laboratory databases and transform the data so that it is understandable for users of the data warehouse.

The laboratory databases aren't all built with a same model and a same DBMS : the main problem was to extract and define the transformation rules necessary to populate the target database with information that is meaningful for data warehouse users.

The integration of data from operational sources is not always easy. As any other data warehouse, our warehouse receives its data from many sources. Each of they require SQL queries to retrieve and move the data from its original location into the warehouse. The Datastage software extracts data and move part of the desired information into the data warehouse.

Some tables have been built for the data warehouse itself : these tables mainly concern the entities "book" and "publisher" and the entities "laboratory" and "team".

Authors and contract researchers are employees of the university : data of this table come from the employee database.

Production table informations yet exit in the different data bases of laboratories. The information of the data warehouse is extracted from this different databases for this particular table.

In the same manner, contract organisation yet exists in the accounting database (in France universities mainly use accounting software named "NABUCO"). In the associated database, all informations concerning contracts are present.

Some informations are not represented in the same way in databases and the data warehouse. For instance, the table "production" cannot be filled without some processes on the data : in laboratory database there is only one information for a given paper and this information is represented as a text, for example the reference [VAS99] P. Vassiliadis, T. Sellis: A Survey of Logical Models for OLAP Databases SIGMOD, a represent by an unique attribute.

It is then necessary to analyse this text in order to found :

The author's name ; first part of the text must be identified and compare with the list of researcher of the laboratory (some authors can not belong to the lab) ;

The rank of authors which belong to the lab, the book and the publisher.

A special software has been created to do this research. In the future, the use of XML will be a good solution to improve this analyse.

3.6 The use of the system

Many queries can be asked using this system. We can give some example of possible queries :

Give all the papers of (a researcher, a team, a laboratory) ;

What is during a given time the average number of publications in reviews for a researcher which belongs to academic staff ;

What are the personal costs of a given team ?

How many papers were published by people working on a given contract.

Appendix

Data warehousing systems have become a key component of information technology architecture. In this paper we presented an architecture of a data warehouse for research centre activity. In the future, the growing importance of XML documents as a means to represent data in the World Wide Web will allows us to extract information from these documents and directly fill the data warehouse.

References

[1] W.H. Inmon: Building the Data Warehouse. ed. Wiley. 1996

[2] R. Kimball: The Data Warehouse toolkit. Wiley Computer Publishing. 1996

[3] P. Vassiliadis, T. Sellis: A Survey of Logical Models for OLAP Databases. SIGMOD Record 28(4). 1999

[4] J. Widom: Research Problems in Data Warehousing. Proceedings of the 1995 International Conference on Information and Knowledge

[5] S. Chaudhuri, U. Dayal: An overview of Data Warehousing and OLAP technology. ACM SIGMOD Record 26 (1). 1997

[6] http://www.olapcouncil.org/

[7] http://www.informix.com/