A National Data Warehouse Project for French Universities

Jean-François Desnos

Université de Grenoble and Agence de Modernisation des Universités, Paris, France

1 Introduction.

The French national Agence de Modernisation des Universités, (AMUE), is the Information Technology Consortium which has as members most of the French Universities, (about one hundred and ten). AMUE provides large management software applications to its members, i.e. student, financial and personnel systems; all of them designed with Oracle client/server technology. AMUE also acts as a consultant for its members in the fields of information technology, university management, and professional development. The Data Warehouse project is one aspect of the management improvement process conducted by AMUE.

Present applications are heterogeneous: the student system has been designed by AMUE and developed by a software company and the financial system has been bought on the market and adapted. The personnel system has been done entirely by AMUE.

Because of this heterogeneity, it is presently difficult, and even impossible, to present reports that cross-reference information coming from several data bases. For example, a report mixing student, staff, and financial data is not really available automatically per request.

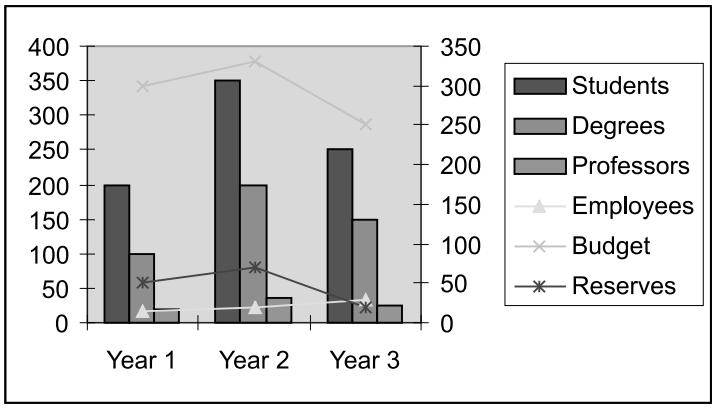
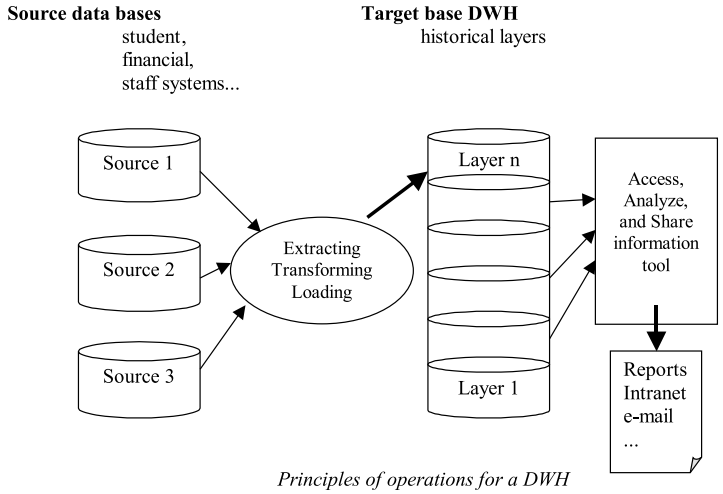


Figure 1: Some key-numbers for a given field of studies

The production data bases are continuously developing and are organized for transactional use: register a student, print a money order, or pay a new employee, for example, and not to extract data sets for an executive information system.

A data warehouse (DWH) provides this last objective. A DWH is built by extracting data from source data bases, verifying and transforming them, and then loading a target data base which becomes the DWH. This process of extraction, transformation and loading, done periodically, provides historical layers of data, which are snapshots of the institutional information system. This DWH is designed principally to edit reports on paper, the Intranet, or by e-mail.

A data warehouse is a set of snapshots of the institution's information system. It is designed to provide on demand indicators to all the concerned individuals of the university.



Principles of operations for a DWH

For French universities, the source data bases are the student, financial, staff systems, all of them provided nationally, plus all the local applications, and also data coming from outside (for comparative reports for example). The DWH has to take into account the complexity, variety and evolutions of the sources data bases and the unsteady quality of source data.

2 The steps of the project.

The project schedule includes the following steps:

2.1 Software specifications: three "scenarios" or classes of indicators have been defined by a working committee, including university presidents, general secretaries, statisticians and information technology managers from nine institutions. The specifications were validated in march 2000.

2.2 Building of a meta-dictionary. A data warehouse meta-dictionary describes the pieces of data extracted from source databases, how the information is treated, and then transfers to the target. This very important step of DWH design was validated by the steering committee in august 2000.

2.3 The development process itself is comprised of:

2.3.1 The target database design. As most of the sources are Oracle 7 based databases, we chose an Oracle 8i standard relational technology.

2.3.2 Programming consists of two kinds of procedures:

- The Extraction, Transformation and Loading, (ETL), from sources to target procedure. To ensure good productivity and maintenance of the software, we chose to use "Data Stage Tool," from Informix, rather than a full in-house software development. After a two-month comparative study, Data Stage has been selected, versus Genio from Hummingbird. We also looked at Oracle Warehouse Builder, and Decision Base from Computer Associates, and Sunopsis.

- Business Objects R5 for target extractions, data analysis (slice and dice), and presentation.

2.4 Delivery of the first release to pilot sites in November 2000.

The five pilot universities are Jules Verne, (Amiens), P. et M. Curie, (Paris), Louis Pasteur, (Strasbourg), Rennes 1, and Versailles.

Four of them: Amiens, Rennes, Strasbourg, and Versailles, use the same student, (Apogée), financial, (Nabuco), staff (Harpège), and payroll (Paye) systems. Major sources thus become the same, but of course, other locally built sources exist.

Paris does not run Apogée, and also had the ETL tool, Genio, before the project began. This pilot institution has therefore to adapt ETL procedures to their student source, and to their ETL tool.

A number of technical documents have been delivered with the software:

- Technical recommendations and specifications,
- Design and development documentation,
- Examples of outputs, (about 30),
- An installation methodology.

2.5 In 2001, a technical meeting is scheduled about every month in Paris with the pilot teams. The steering committee meets every two months.

- A minor corrective release was delivered in January.
- A major new release, including evolutions of the target and a second B.O. universe will be ready in April.
- A national seminar will be held in May hosted by Agence de Modernisation to report on this experiment to all interested French universities.
- In July 2001, a CD-ROM including all software and documentation of the project will be available at no charge to all interested universities.

3 The specifications.

The committee, which made the specifications for the pilot issues of the data warehouse, defined three executive information boards, which have been called "scenarios".

3.1 The first scenario is about the evolution of the number of professors with regard to the research and teaching needs. The aim is to help a president's decision about recruitment.

It produces a set of reports on the number of students and teachers per the field of study, and the ages of teachers, (to forecast retirement schedules). The same information is provided regarding laboratories, thesis directors, and researchers.

3.2 The second scenario is a synthetic presentation of the faculties making up the institution.

Human Resources, (academics, associates, staff), and the financial means regarding number of students, number of degrees offered, and the developments over several years are analysed and reported in this scenario.

3.3 The third scenario aims to measure the international attractiveness of the institution. The purpose is to control the institution's policy on international reputation and exchanges.

All students involved in exchange programs with foreign universities, invited professors, and the exchange of scientific and collaborative programs are listed per field and academic year with the budgets involved.

4 The target database.

The target or data warehouse itself is a ROLAP Oracle 8i database, designed the following way:

4.1 Four principal tables: STUDENT, PERSONNEL, BUDGET, and HC, (HC is made of extractions from payroll). Each of these tables gathers information from the various operational applications.

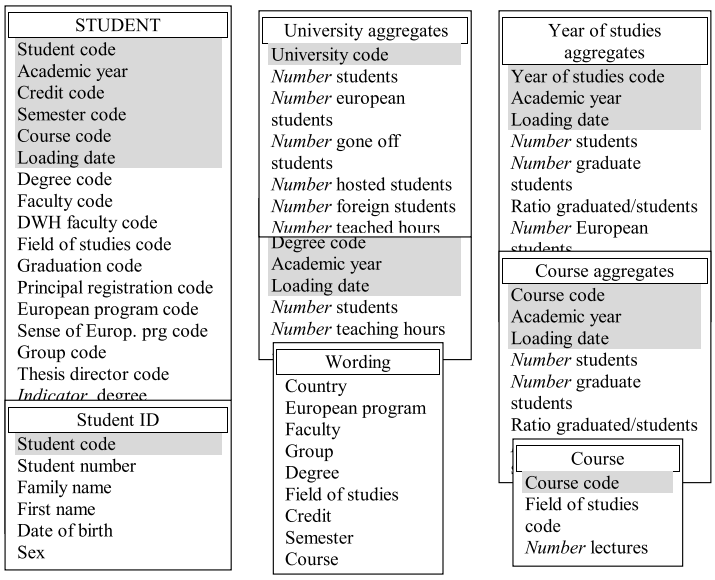
4.2 A number of wording and aggregate tables. Columns of aggregate tables are calculated during the loading process: for example the number of professors in a given field.

4.3 Two "service" tables:

- The "structure of the university" table. This table gives a unique code to each faculty member, and each school of the university. This is necessary because all of the operational source application has its own code and wording.
- The "observation dates" table. This table will permit the extraction of Business Objects in different layers for each of the source databases. This means that it can be edited as a report based on: the snapshot on 15/01/2001 of the student system, and the snapshot on 01/01/2001 of the financial system, etc...

In these tables the value of an indicator is a number (derived from a boolean) the value of which is 0 (false) or 1 (true). Indicators will simplify the scorings with Business Objects.

The STUDENT and its subsidiary tables are figured thereunder as an example, avoiding specific French system data.



5 Metadata.

Metadata includes definitions on used items, (aggregates, fact tables), of the target data base, on the data itself, and the way they are extracted, calculated, and inserted in target tables and columns. Our metadata repository being a 55 page document, is listed here with only a few general notions.

Academic year: it's a key concept of the DWH. For most of extracted data, the reference year is the academic year, except for the financial data, where it is the calendar year. The existence of two reference periods is one of the difficulties of the DWH building and querying.

Course: the educational organization is tree-structured. Only terminal elements of the tree-structure are used for calculations.

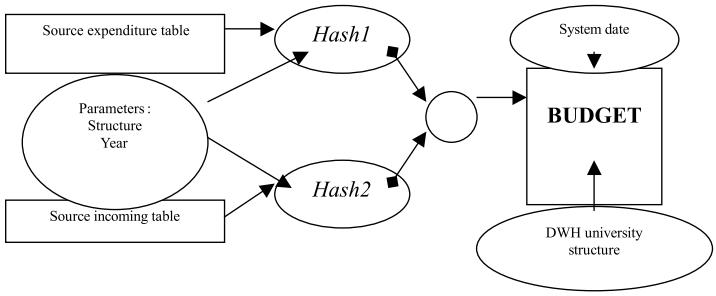
Faculty, school: some of them do not appear in all the source application. The structure of the institution and the codes can also be different. A common "institution structure" table, with source and structure links, has been added to the DWH tables.

Teaching units: the payment of teaching hours for an associate professor depends on what is done: lectures,tutorials,labs, and on the teacher's rank.

6 Data Stage job example: loading the budget in the BUDGET target table.

Job parameters are (1) university structure code and (2) a set of calendar years.

- Expenditure data is extracted from the financial source database. Data is written in a hash file for performance reasons.
- Incoming is extracted from the financial source database. Data is written in a second hash file.
- The BUDGET target table is updated from hash1 and hash2, with the loading date and the university structure code (expenditure and incoming do not necessarily correspond).



Loading DWH BUDGET table

Why use a commercial ETL tool ?

ETL (extraction transformation loading) can be developed using SQL and C for example, and not a specific tool, which is expensive (15,000 to 50,000 Euro).

But such an ETL software allows:

- A faster development,
- More reliable corrections and evolutions,
- A better adaptation to changes in sources and target bases,
- An adaptability to a changing information system.
- When operating, a loading of layers ("snapshots") by a planning tool.

7 Target queries.

With the present release of the project (March 2001) two Business Objects universes are provided.

Business Objects, (BO), is a commercial software application used to extract, analyze, and present data discovered in a data base, (data mining). BO is designed to make these operations easier. A BO query extracts data, then presents it in sheets which can be dynamic, (multidimensional "slice and dice"), and published on the Intranet.

BO is used by two categories of users:

- IS specialists, who define the universe. In other words, the architecture of data extraction.
- End users, who can either create new queries, or simply set the parameters and run existing queries.

8 Conclusion.

This project is a national attempt to give each French universities a starting core model for the future development of its own Decision Support System. This was feasible because French universities organised, (in 1992), a software consortium which provides most of the members their main administrative systems. The pilot universities will test the prototype, and adapt it with the project team, until July 2001. A free delivery to volunteer universities will take place from the end of the year 2001. This project is part of a general modernization of management processes started in the universities under the auspices of the Agence de Modernisation des Universités.

9 Acknowledgements.

Many thanks to:

- the modernisation of management process group at Agence de Modernisation des Universités, leaded by C. Charrel and S. Rochas (AMUE), for fruitful meetings,
- the technical teams of pilot sites for their help in implementing and discussing the model: E. Cravoisier (Amiens), B. Perrigault, A. Routeau (Rennes-I), F. Cadé, O. Raunet, V. Vaillant (Strasbourg), R. Casteloot, O. Morelle, J.-B. Nataf (Paris-VI), D. Fiquet, N. Courtay, R. Rivoire (Versailles),

Grateful thanks to the Grenoble AMUE design and development team: Marie-Hélène Glénat, project assistant leader, and Nicolas Maume, my former student.

References

[1] Building the Data Warehouse (W.H. Inmon)
[2] The Datwarehouse Toolkit (Kimball/Wiley)
[3] The Datwarehouse Lifecycle Toolkit (Kimball/Wiley)
[4] DataWarehousing with Oracle (Yazdani-Wong/Prentice Hall)
[5] Atre's Road Map for DW/DM Implementation written (Shaku Atre)
[6] Olap Solutions (Thomsen/Wiley)
[7] Data Warehousing and Data Mining for Telecommunications (Rob Mattison/Artech House)
[8] The Data Warehouse Challenge (Michael Brackett)
[9] Oracle 8 Data Warehousing (Dodge, Gorman)
[10] Data warehousing in the real world : a practical guide for building decision support systems (Anahory, Dennis Murray)