An Academic Data Warehouse

Carlo DELL'AQUILA, Francesco DI TRIA, Ezio LEFONS, and Filippo TANGORRA
Dipartimento di Informatica
Università di Bari
via Orabona 4, 70125 Bari
ITALY

Abstract: - There are several benefits that can be reached by developing an academic data warehouse as providing a centralized source of information accessible across different academic units to quickly analyze problems and get satisfactory solutions, supplying the data necessary for developing the Institution's strategic plan, and enabling administrator to make better business decisions based on historical data available in legacy databases. The paper describes the architecture of an academic data warehouse. Examples of analytic reporting are also reported.

Key-Words: - Data warehouse, Data mart, OLAP, Academic application

1. Introduction

Data warehouses (DWs) are the most significant component of strategic decision making for business. Since 1970 new approach to analyze business data has become important for those companies, as banks, financial services, or chains of supermarkets, for which the customer satisfaction is the key of success. However, in the early years, the costs for the development of a data warehouse were very expensive. Only recently because of the lowering of the cost for developing and maintaining a data warehouse, these databases designed to support managerial decision making have become functional tools to use as repository of information [1-4]. Also Universities, that until this moment were almost absent in the list of the major users of data warehouses, have accepted to take advantage of developing a decision support system in an academic environment [5]. In fact, nowadays, we can consider the management of a University as critical as the management of a big business company, because the factors affecting an optimal management of a University are the same involved in the business processes.

Typical objectives affecting the management of a University are:

- offering a better quality of the instruction;
- managing employees and human resources;
- managing economic-financial institutions;
- avoiding wastes.

There are several environmental factors that have to encourage academic institutional leaders to adopt an academic data warehouse. These factors include not only decreases in governmental financial support, faculty supplies and research founding, but also increases in student tuition, competition, faculty salaries, faculty support and expectations from students, parents, and employers. Each of these factors generates informational drivers for the development of an academic data warehouse. One driver is represented by the necessity to follow the pace of change affecting business companies; this driver obligates academic institutions to gather information to support strategies and processes that address changes. Another driver is to provide a centralized repository that represents a centralized tool for all the decision makers to control global resource allocation and use.

Given these information drivers, there are several benefits that can be reached by developing an academic data warehouse. For example,

(a) providing a centralized source of information accessible across different academic units to quickly analyze problems and get satisfactory solutions,

(b) supplying the data necessary for developing the Institution's strategic plan, and

(c) enabling administrator to timely make better business decisions based on historical data available in different data stores.

This paper presents the architecture, and design of an academic data warehouse supporting the decisional and analytical activities regarding the three major components in the university context: didactics, research, and management.

2. The National Context

In our national context, the most significant experience about developing applications related to the various university management needs has been made by CINECA [9]. CINECA is a non-profit Interuniversity Consortium, made up of 28 Italian universities. Due to its nature, the consortium follows with great attention the national normative evolution continuously adapting the released applications or developing new ones.

The main developed applications aim to manage:

- the legal-economic career and the wage of the academic and technical personnel;
- the function and the activities of the employees;
- the career of the students of the Athenaeum and the didactic programming;
- the economic and financial resources.

The consortium proposes also a data warehouse for analytic activities.

Nevertheless, each University adopts only the services it chooses from those developed by the consortium. Moreover, each University has own legacy databases of historical data. It follows the need of data integration in order to access all these information resources mainly for analytic purposes.

Also our University meets the previously described conditions. Therefore, we decided to develop a specific data warehouse integrating all the present and historical data resources.

3. The Academic Data Warehouse Architecture

Business Intelligence consists of applications and technologies that help companies to have a wide knowledge about their own business performances. A Business Intelligence System in the University context has a wide knowledge about the performances on students, teaching staff, and Didactics. A University data warehouse is designed to provide a valid tool that satisfies the following needs:

- a unique system of analysis and reporting for the supervisory staff of the Athenaeum and for the single organizational and administrative structures, such as departments or secretariats for the students;
- a system that supplies in real time data to information external agencies.

Figure 1 shows the University data warehouse architecture structured on the typical multi-level layout.

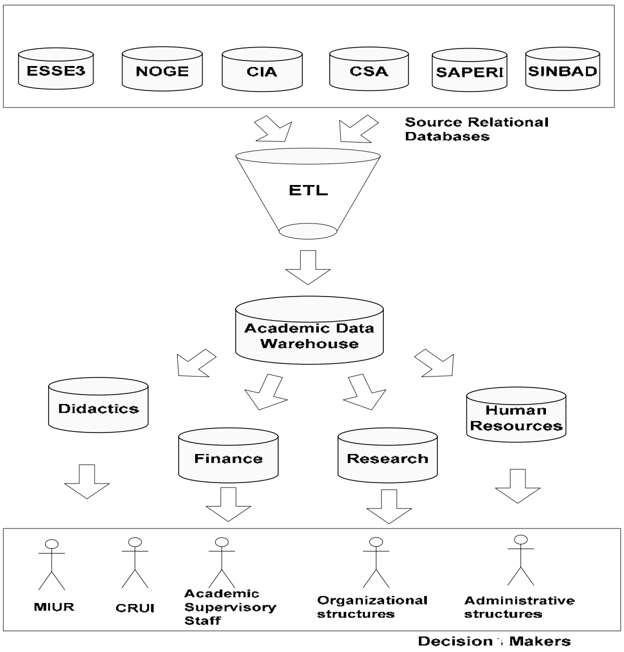

Fig. 1. The academic data warehouse architecture.

3.1. The Source databases

Source databases contain transactional data. Figure 1 shows six source databases.

ESSE3 (Secretary and Services for Students) is the new database that supports all the didactic curricula, and administrative processes and services to the students with the respect of the didactic autonomy of the University.

NOGE (NOt manaGEd) is a secondary old database that stores residual historical data about students enrolled before the ESSE3 introduction.

CIA (Athenaeum Integrated Accounting) is the integrated financial management system that considers the University as a business company that distributes specialized services (Research and Didactics, for example).

CSA (Careers and Wages of Athenaeum) takes care of the legal and economic management of the university personnel.

SAPERI is the database of the scientific research competence of the University. It includes also publications and patents of researchers. These data concur, for example, to construct the athenaeum yearbook.

SINBAD is the system for the management of the athenaeum research projects.

3.2. The ETL process

The ETL process loads data from source databases into target tables of the data warehouse. This process requires a deep knowledge of the schema of the source databases, in order to map every field of the source tables with the opportune field of the target tables, and store data. In fact, our old source databases (*e.g.,* NOGE database) contain dirty data due to input errors and the lack of controls by the software used for storing data. These errors frequently are null values or typos. However, severe errors are consistency errors or the presence of duplicated records.

Therefore, in the ETL process, there occur two kinds of problems. The first one arises in populating dimensional data cubes. In fact, typical errors coming out when loading data into a dimension table are the violation of the primary key constraint. The second kind of problems regards the data mapping to multidimensional structures, in that foreign key constraint violations occur frequently.

3.3. Data warehouse

After the ETL phase, the data warehouse contains cleaned historical and integrated data. The data warehouse is composed of a set of data marts to model the following academic functional areas:

a. **Didactics**. This data mart contains data about the career of the students of the Athenaeum. Moreover, there are information on the University formation offer structured in Faculties and Degree Courses.

b. **Finance**. This data mart is developed to run twofold analyses: (a) the analysis of financial documents, and (b) the analysis of general and analytic economic movements.

c. **Research**. The research data mart contains awarded research projects and applications for research grants. It also contains data on components and location of every research project.

d. **Human Resource**. The model adopted for this functional area allows to investigate on the legal-economic careers and wages of the academic personnel. Moreover, it allows to extract information related to the functions, activities, and location of the academic, administrative, and technical personnel.

3.4. OLAP layer

The data warehouse supports OLAP queries producing reports for managers and decision makers. In Figure 1, there are shown the decision makers. There are internal or national decisional agencies.

a. **Academic Supervisory Staff**. There are two principal Academic Supervisory Staff: The Academic Senate and the Administration Council. The Academic Senate is the governing body in matter of programming the development of the Athenaeum and the coordination of Didactics and Research. It approves the criteria for the distribution of the financings among the Research Structures. Moreover, it determines the criteria for the evaluation of the didactic activities and estimates the effectiveness by analyzing the report produced by the Evaluation Team. This is a partially elective independent team, named by the University Rector, with the function to verify periodically the operating efficiency of the didactic structures, research structures and structures for the technical-administrative management.

   The Administration Council deliberates and supervises the administrative, financial, and economic-patrimonial management of the Athenaeum. In particular, the Council deliberates about the performance of the criteria for the distribution of the financial resources among institutions and the technical and administrative staffs of the University.

b. **Organizational structures**. Faculties are the fundamental structures that organize and coordinate the Didactic activities. In University, the management of the Research activities is entrusted to the Departments. The Departments are the organizational structures that collect teachers and researchers coming from several Faculties, but joined by the same scientific interests and research methodologies. The Departments collaborate with the Faculties for the realization of the Didactic activities.

c. **Administrative structures**. These structures are the student secretariats and the data elaboration centres, whose tasks are the production of data for the national “Alma Laurea” registry of the graduate students and the realization of documents, statements and other information prospects to support the decisional processes.

d. **MIUR**. The national committee for the evaluation of the university system is the MIUR institutional team, whose tasks are: to establish the general criteria for the evaluation of the activities of the university; to predispose the annual report on the evaluation of the university system; to promote the experimentation, application, and spread of methodologies and evaluation tasks; to determine the nature of the information and data that the athenaeum evaluation team must communicate; to predispose studies and documentation on the state of the university instruction, the compliance with the study right, and the accesses to the university courses of study.

e. **CRUI**. The CRUI is the Association of the Rectors of the Italian Universities. It was born in 1963 as a private association of the Rectors and, in short time, it has acquired a recognized institutional role and a concrete ability to influence the development of the university system through an intense activity of study and experimentation. CRUI centralizes its own evaluation activity in particular on the Didactics and Research areas, develops and proposes methodologies and evaluation criteria for athenaeums, and degree courses, finalized to the improvement of the quality of the Italian university system.

4. The Didactics Data Mart

The Didactics data mart uses mainly two source databases to load and refresh data: ESSE3 is the main database, and NOGE the second one. Figure 2 reports detailed logic model of the Didactics data mart for the academic data warehouse.

The Didactics data mart contains the following fact tables: enrolment, tax, examination, and degree. All these fact tables have three dimension tables in common: student dimension, degree course dimension and time dimension. These are the basic dimensions, because they represent the minimum of information to express «who, where and when» aggregation levels.

The **enrolment** fact table has five dimension tables. The additional dimensions are: residence, that allows demographic or geographic aggregation, and kind of enrolment, that allow administrative aggregation. This table has no measures and its function is to store the enrolment to a course of study by the student.

The **tax** fact table has only the student, time and degree course dimensions and it has the amount field as measure. Its function is to control the payment of the taxes by the student.

The **examination** fact table has four dimension tables; the additional dimension is represented by the teaching course, that allows didactic aggregation. It has two fields: the first field is the mark, that represents the fundamental measure; the

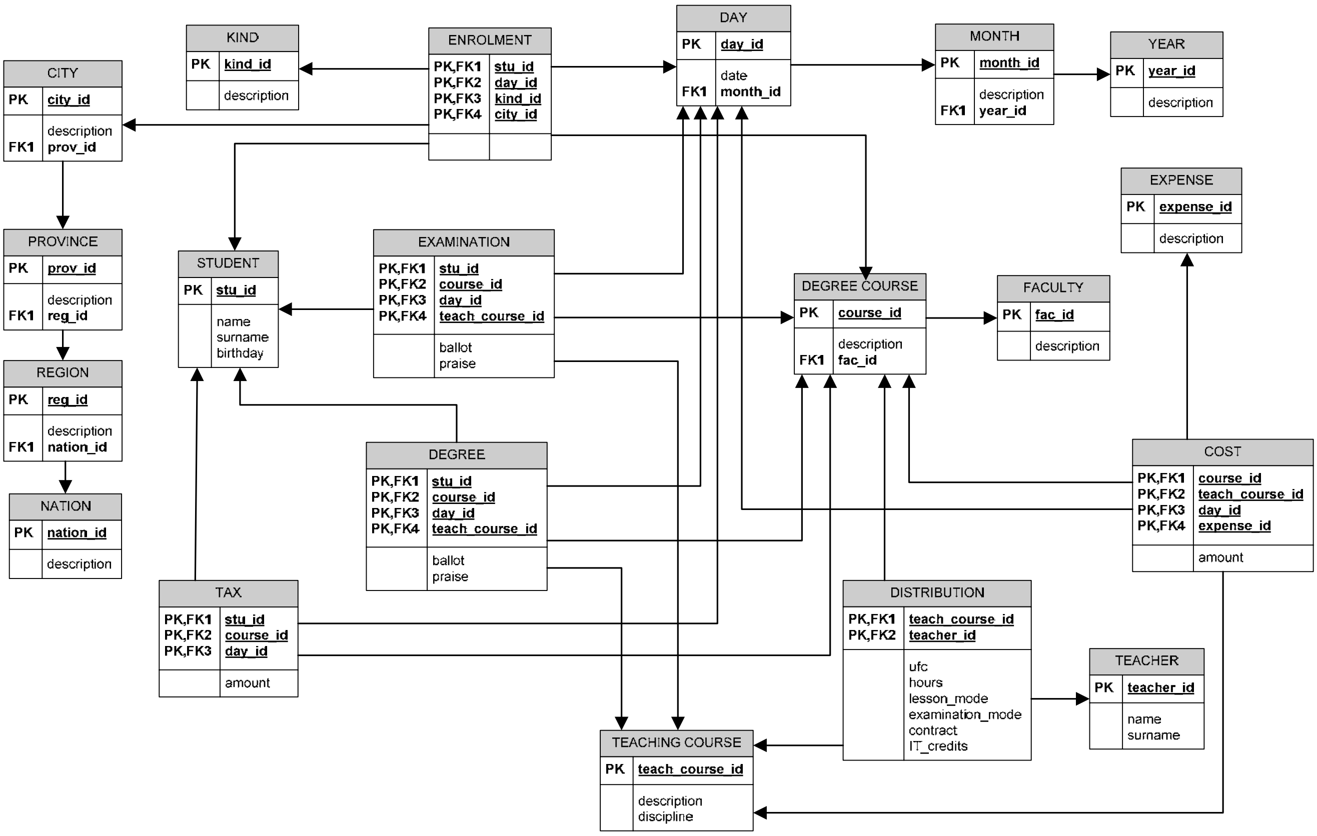

Fig 2. The Didactics data mart.

second is the *cum laude* field, that is a simple Boolean field.

The **degree** fact table has four dimension tables too; it has the same additional dimension owned by the examination fact table and it has the same measure and fields: mark and *cum laude*.

Moreover, in this data mart there are further fact tables to allow analyses and statistics on the didactic offer of the University.

The **distribution** fact table indicate the list of teaching course for each degree course of study. The information includes the number of teaching hours, the number of university formative credit (UFC), the kind of lesson, the kind of examination, and also the teacher for each teaching course.

The **cost** fact table is relative to the annual costs supported by the University for the management of each teaching course, and totally for each degree course per academic year. It contains also information on the teacher's cost, when the teacher is not enrolled in the University teacher's staff.

To obtain aggregate results at a different level of granularity, some dimensions are organized in dimensional hierarchy. In particular, the degree course is a two-level dimensional hierarchy: course and faculty, for allowing the aggregate measures (for example, the count of graduate student) at the degree study level or at the faculty level. The residence is four-level dimensional hierarchy for aggregate measures at the city, province, region, nation levels for analyzing data aggregation referring to different geographic contexts. Finally, the time is a three-level dimensional hierarchy including the day, month and year levels, for summing data respectively by day, by month or by year. All other dimensions of the didactic data mart are one-level hierarchy

5. Reporting applications

Data analyses with OLAP and data mining techniques are used to achieve reports and responses to complex queries. For example, referring to the Didactics data mart, objectives of investigations can regard information on the student, forecasts of the formation trend of the Athenaeum, such as:

- Monitoring the incoming and outgoing flows of the students in the University: the count of matriculated students grouped by academic year, the count of enrolled students grouped by academic year and course year (distinguished in regularly enrolled students or outside run students), the count of graduated students by session and degree course, the count of successful examinations and the average mark by academic year and teaching course, the average count of successful examinations by course year.
- Monitoring the didactic workload of the teaching staff: number of hours for teacher and for didactic activity, number of presences in commission of profit or degree examination with various roles, number of reported degree theses.
- Monitoring the financial trends of the student taxes.
- Monitoring the needs of a teaching subject matter (Informatics, Mathematics, Foreign Languages, …) in the didactic offer of the University.

The system can produce traditional statistics as those reported in Tables 1 and 2. The report in Table 1 considers the students enrolled between years 2000 and 2005 grouped by academic year and region. The report shows the regions that represent the most important affluence centres for our University. We observe that, because residence dimension is a four-level dimensional hierarchy, the same analysis with a roll up operation can produce a coarser grain result summing data for nation or, with the opposite drill down operation, provide a finer-grained view considering the number of students at the province or at the city levels.

The report in Table 2 groups the same data by academic year and university Faculty. In this case, because degree course is a two level dimensional hierarchy, the same analyses with a drilldown operation, provide a detailed map showing the counts of students grouped by degree course.

Complex analyses are accomplished as in the Table 3, that reports the presence of the Informatics credits in various university degree courses. The aim of this analysis consists of listing all Informatics teaching needs in the University degree courses,

REGION	2000	2001	2002	2003	2004	2005
Puglia	64396	65037	70529	68542	60674	60499
Basilicata	3264	2945	3088	2824	2460	2691
Calabria	813	739	778	778	649	719
GREECE	306	250	237	219	155	102
Lombardia	135	106	104			
Lazio	129		119	103		
Campania	123	129	208	260	175	332
Molise			221	136	106	
Sicilia			114	127		163

Table 1. Count of enrolled students grouped by Academic year and Region.

FACULTY	2000	2001	2002	2003	2004	2005
Law	15222	13092	12634	11408	8619	9026
Economics	9935	9797	10235	9254	7496	7904
Educational Sciences	8920	9814	11187	12346	13064	11963
Mathematics, Physics and Natural Sciences	6806	8525	9821	8907	6795	6622
Medicine and Surgery	5659	5427	6345	7075	7464	8080
Arts and Philosophy	5409	4792	5675	5288	4506	4920
Political Sciences	4732	4388	4598	4268	3092	3417
Pharmacy	3786	4242	3550	3700	3984	4430
Foreign Languages and Literatures	3318	3532	4304	4364	3841	3833
Law (Taranto city)	1816	2148	2806	2862	2278	2164
Veterinary Medicine	1431	1332	1592	1744	1835	1799
Agricultural Sciences	1296	1079	1530	1347	938	821
Economics (Taranto city)	901	801	745	689	584	693

Table 2. Count of enrolled students grouped by Academic year and Faculty.

DEGREE COURSE	TEACHING COURSE	DISCIPLINE SECTOR	LECTURE UFC	LECTURE HOURS	LAB UFC	LABORATORY	TEACHER			ALLOWED I.T. CERTIFICATION	
							STRUC-TURED	DISCIPLINE SECTOR	UNDER CONTRACT	ECDL	OTHER
Physics	Informatics Fundamentals	ING-INF/05	6	48	2	base	y	ING-INF/05			
Physics	Programming Languages	ING-INF/05	1	8	2	specialist	y	INF/01			
Cultural Heritage	Informatics Applications	INF/01	6			none			y		
Cultural Heritage	Informatics	INF/01	4	32	2	base			y		
Mathematics	Informatics	INF/01	7	42	2	base					
Bio-sanitary Science	Informatics	INF/01			3	base				y	MS Certif.
…	…	…	…	…	…	…	…	…	…	…	…

Table 3. Informatics Teaching Courses of Degree Curricula.

showing the UFC credits, teachers, and – if allowed – equivalent I.T. certifications. Analyzing this report, the Academic Senate can obtain indicators on the quality level and teaching efforts (in term of teacher's and teaching costs) about the Informatics teaching in the University.

6. Conclusion

The paper summarizes the experience in designing and modelling an academic data warehouse. Existing facilities and databases affect the chosen data warehouse, that brings them together to support decisional activities leading the whole university environment, including administrators, faculties and students. The choice to develop a dedicated system is mainly forced by the peculiar information type that defines the basic information in data warehouse widely different from institution to institution.

Future work will provide the extension of the system with a high-performance layer for describing and managing data profiles in the warehouse [10]. This will be done in order to provide approximate query processing for OLAP applications that allows more speed analytical query.

References:

[1] W. H. Immon, *Building the Data Warehouse*, John Wiley & Sons, 1996.

[2] S. Chaundhuri, U. Dayal, and V. Ganti, Database technology for decision support systems, *IEEE Computer*, Vol. 34, No 12, 2001, pp 48-55.

[5] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, *Fundamentals of Data Warehouses*, Springer-Verlag, 2003.

[4] R. Kimball and M. Ross, *The Data Warehouse Toolkit*, 2nd edition, John Wiley & Sons, 2002.

[5] D. Wierschem, J. McMillen, and R. McBroom, What Academia Can Gain from Building a Data Warehouse, *EDUCAUSE Quarterly*, Vol. 26, No. 1, 2003, pp 41-46.

[6] G.L. Donhardt and D.M. Keel, The Analytical Data Warehouse: Empowering Institutional Decision Makers, *EDUCAUSE Quarterly*, Vol. 24, No. 4, 2001, pp 56-58.

[7] M.C. Lin, University Data Warehouse Design Issues: Case Study, *Proc. of the 2001 American Society for Engineering Education Annual Conference & Exposition,* 1-9.

[8] C. Fernandes and M. Whalen, Data Warehousing from the Web, *Proc. of the 2004 American Society for Engineering Education Annual Conference & Exposition*, 1-11.

[9] www.cineca.it .

[10] C. dell'Aquila, E. Lefons, and F. Tangorra, Decisional portal using approximate query processing, *WSEAS Transactions on Computers*, Vol. 2, No 2, 2003, pp 486-492.