Business Intelligence Solution for University Management

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

Abstract: - Traditional users of data warehouses were banks, financial services, or chains of supermarkets. Instead, Institutional Organizations (e.g. Academies) in the past did not use the large amount of transactional data for strategic decision making. The optimal management of a University can now be considered as critical as the management of a big enterprise. In fact, the factors affecting the management of a University are the same involved in the business processes. The paper describes a proposal of the architecture of a Business Intelligence system and the flow of data processing for our University.

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

1 Introduction

Business Intelligence is an activity based on software technologies whose aims are to support the management and the processing of the data in an Information System. This data processing, that ends with the production of information usable by managers and business decision makers, starts with collecting and storing large volume of historical and heterogeneous data in a central repository: the data warehouse (DW). For this reason, DWs are the most significant component of strategic decision making for business. In the last years, a 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 [6-8]. For example,

1) providing a centralized source of information accessible across different academic units to quickly analyze problems and get satisfactory solutions,
2) supplying the data necessary for developing the Institution's strategic plan, and
3) 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.

In this paper, Section 2 introduces the current status of the systems supporting the development of Business Intelligence applications at the national level. Section 3 describes our Business Intelligence system, that comprises the architecture, the source databases, the designed data warehouse, and the OLAP layer supporting academic decision makers. Section 4 shows the logical schema of the data warehouse. Section 5 presents some typical Business Intelligence applications developed with our system. Section 6 reports conclusions and future works.

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 Business Intelligence 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.

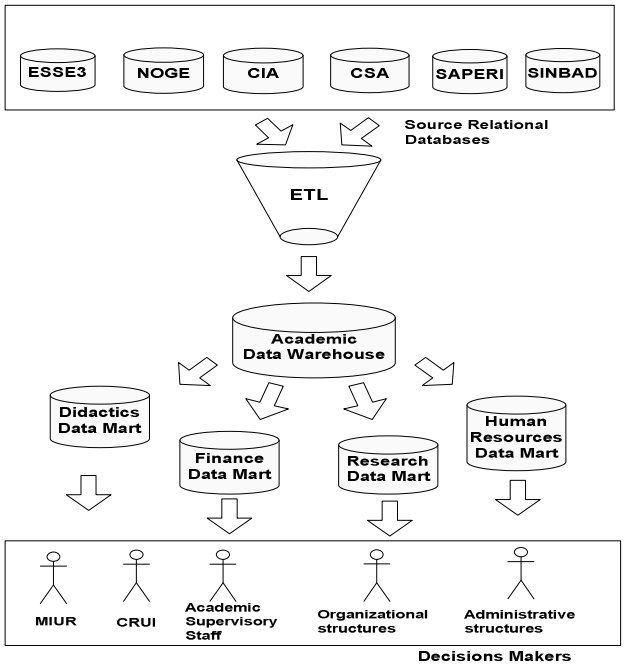

Fig. 1. The academic data warehouse architecture.

3.1 The Source databases

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

ESSE3 (Secretary and Services for Students) is the current 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 legacy 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 Data warehouse

After the ETL (*Extraction Trasformation Loading*) phase, the data warehouse contains cleaned historical and integrated data.

Given the high complexity of the data warehouse, at the present moment, we have started to design and to implement the single data marts that represent the departmental databases, on which we will construct the whole data warehouse as future work.

The current data marts are able to model the following academic departmental 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.

3.3 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

This Data Mart has been designed through the integration of the logical schemas of the two transactional databases: (a) ESSE3, the current database that supports all the didactic curricula, and administrative processes and services to the students; (b) NOGE the secondary database that stores residual historical data about students enrolled before the ESSE3 introduction.

NOGE and ESSE3 represent also the repository of data used to feed the Data Mart. The *Extraction Transformation Loading* (ETL) process is the step in which data are loaded from the source table and stored in the tables of destination. There are significant differences between the two databases, that made very difficult the process of data integration. These differences regards not only the cardinality of the tables but even the representation of the information.

ESSE3 is a database where the cardinality of tables is in the order of millions of tuples. The only errors can be found in string fields and generally consist of typos. In NOGE, there are tables with thousands of records, but, in spite of this, the database contains very dirty data. Typical errors are null values and typos, due to the lack of control by the software in the data entry phase and to the field type chosen to format data. For example, a date is usually stored in a string field in *yyyymmdd* format, allowing the user to insert partial or inconsistent dates and making impossible DBMS control. Moreover, the most serious problem regards the presence of redundant and duplicate records, often inconsistent among them. For example, the examination of a student on a specific study course may be reported two times with the same date and the same mark, but with different values in the *cum laude* field.

Obviously, the two databases adopt different data entry standards. For example, the tables that store data about the Departments of the Athenaeum can differ significantly in the name used for the structure. It is possible to find indistinctly strings as "Informatics Department" or "Department of Informatics".

All these problems have been faced in data cleaning, that is a sub-process of the ETL process in which it is necessary to solve instance-level problems relative to actual data contents which are not visible at the logical schema level. In general, misspellings, typos, abbreviations, and conflicting representation errors can be identified and corrected by a spell checking, based on dictionary lookup that includes synonyms, and by transforming all the data in a standardized form. Duplicate record eliminations require to identify similar records that refer the same real world object, by "fuzzy matching" techniques based on matching rules [10].

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, degree course and time. 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 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*.

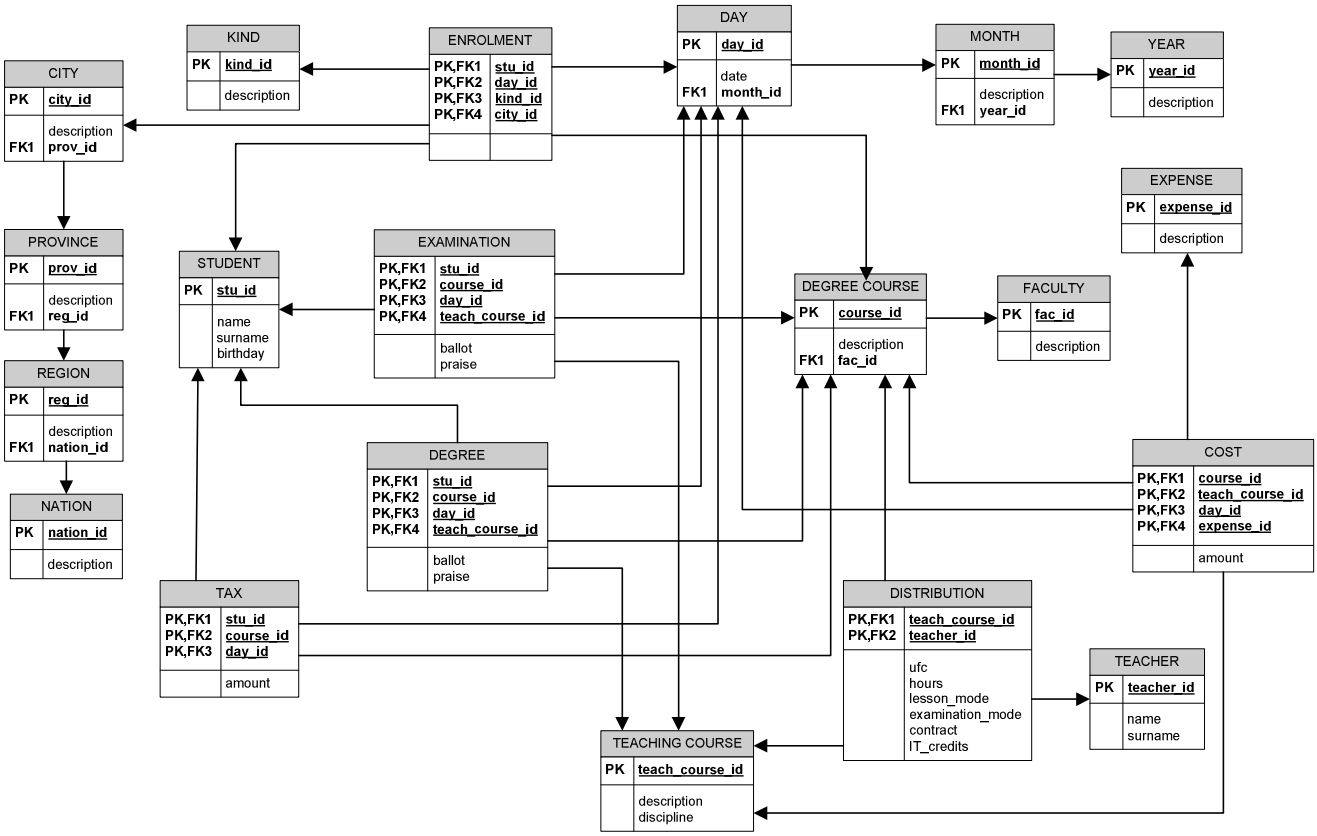

Fig 2. The Didactics data mart.

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 grouping 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.

This report shows the most populated Faculties of the University of Bari. In particular, the information it is possible to extract is that in 2005, the last considered year, the Educational Science Faculty had the most affluence with the 18.22% of students, while the Economics Faculty, localized in Taranto city, had an irrelevant number of students with the value of 1.06%.

Traditionally, a large number of students have always chosen the Law Faculty, that was, for social and historical reasons, the most populated Faculty of the University. After 2004, this Faculty have loosed almost the 50% of its students, that are migrated to Educational Science, Medicine, and Surgery Faculties. The rest of the others Faculties does not show significant difference during the five years.

We observe that, because Faculty dimension is a father dimension for Degree Course, the same analysis with a drill down operation, provide a finer-grained view that allows to consider the percentage of students for each Degree Course of all the Faculties.

Complex analyses are accomplished as in the Table 2, 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, 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 overall Informatics teaching in the University.

FACULTY	2000	2001	2002	2003	2004	2005
Edu. Sci.	12.88	14.23	14.91	16.85	20.26	18.22
Law	21.99	18.98	16.84	15.57	13.36	13.74
Med. and Surgery	8.17	7.87	8.46	9.66	11.57	12.30
Economics	14.35	14.20	13.64	12.63	11.62	12.04
Math., Phys. And Nat. Sci.	9.83	12.36	13.09	12.16	10.54	10.08
Arts and Philosophy	7.81	6.95	7.56	7.22	6.99	7.49
Pharmacy	5.47	6.15	4.73	5.05	6.18	6.75
Foreign Lang. and Lit.	4.79	5.12	5.74	5.96	5.96	5.84
Political Sciences	6.84	6.36	6.13	5.83	4.79	5.20
Law (Taranto city)	2.62	3.11	3.74	3.91	3.53	3.30
Veterinary Medicine	2.07	1.93	2.12	2.38	2.85	2.74
Agricultural Sciences	1.87	1.56	2.04	1.84	1.45	1.25
Economics (Taranto city)	1.30	1.16	0.99	0.94	0.91	1.06

Table 1. Percentage of university students, grouped by Faculty from 2000 to 2005.

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 2. Informatics Teaching Courses of Degree Curricula.

6 Conclusion

The paper summarizes the experience in designing and modelling an university 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 [11]. This will be done in order to provide approximate query processing for OLAP applications that allows more speed analytical query.

Acknowledgements:
We tank the CSI (Informatics Service Center) of the University of Bari.

References:

[1] W. H. Immon, *Building the Data Ware-house*, 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.

[3] M. Jarke, M. Lenzerini, Y. Vassiliou, and P. Vassiliadis, *Fundamentals of Data Ware-houses*, 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,* pp. 1-9.

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

[9] CINECA-Consorzio Interuniversitario, Bologna, www.cineca.it.

[10] E. Rahm and H. Hai Do, Data Cleaning: Problems and Current Approaches, *IEEE Bulletin of the Technical Committee on Data Engineering*, Vol. 23, No. 4, 2000, pp. 3-13.

[11] 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.