Challenges in developing a cost-effective data warehouse for a tertiary institution in a developing country

A. Nazir & T. McDonald
Department of Computer Science and Informatics,
University of the Free State, South Africa

Abstract

Higher Education institutions have grown through the years and have developed into large businesses. Even though industry has experienced a dramatic increase in the use of data warehousing techniques, tertiary institutions have been slow to follow suit. Reasons for this can be attributed mainly to the many reported failures and the costs involved there in. Several factors are forcing these institutions to embrace data warehousing technologies. This paper will report on the challenges involved in taking this step and will show that a successful data warehouse can be developed, regardless of the obstacles involved.
Keywords: higher education, data warehousing, student data mart, star schema, developing countries.

1 Introduction

Industry in developed countries is experiencing a dramatic increase in the use of Data Warehousing (DW) techniques. This can, however, in most cases only be achieved at huge cost. According to the Western Michigan University [1] the primary consideration in the development of a DW is cost. Wierschem *et al.* [2] indicates that a DW requires millions of dollars to develop, plus significant hardware and personnel investment. Hammond [3] quotes a Meta Group survey that the average cost for an enterprise warehouse is $3 million. This can be a major obstacle in developing countries.

Wagner *et al.* [4] stated that the budgets of developing countries are not even sufficient to pay for the Knowledge Management (KM) enabling IT architecture.

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060391

Less developed countries like the Philippines and Pakistan spend a smaller percentage of their budgets on IT, as compared to developed countries like the United States and the UK. For example, the Philippines spend 0.8 percent of its budget on IT, compared to the United States which spends 13 percent. From the above statistics there is a great need to find ways for entering DW technology in developing countries with their limited budgets for IT.

In South Africa (SA) a limited number of organizations are using DW and this technology is still emerging. Pioneering organizations like Electricity Companies and Banks set up very large databases for their management and executive information systems well before the warehouse concept was established (Griffiths [5]). DW technology entered SA in 2001 and a number of organizations like telecommunications companies and banks are currently using DW technologies successfully.

Higher Education (HE) has been slow to follow suit [6]. Before the advent of democracy in 1994, the SA government's tertiary education funding policies mirrored apartheid's divisions and the different governance models which it imposed on the HE system (Bunting [7]). For the new government that came into power in 1994, the focus was to address the imbalances of the past, especially health, housing and primary education. The result was that the subsidies allocated to universities (primary source of income) have drastically been cut. Most of universities that still survive today had to go through a period of tough rationalizations. The bad part of all this is that universities in SA now run on limited budgets. The good part is that they have grown through the years and educational institutions have developed into large businesses in and of themselves (Desruisseaux [8]). This change has resulted in a more business-like management of these institutions as well (Lazerson *et al.* [9]).

It is clear from the above that a number of factors are forcing universities in the direction of DW. Wierschem *et al.* [2] states that the environmental factors that encourage academic institutions to investigate DW options are decreases in governmental financial support, faculty supplies, and research funding, and increases in student tuitions, competition, faculty salaries, faculty support and the expectations from students, parents and employers. Add to the above list the fact that universities must nowadays compete as businesses in order to survive.

To enter this new world of DW, tertiary institutions must face a number of challenges. Cost is one of the challenges. Another one is that many of the current OLTP systems lack data integrity and errors abound. Still another challenge is that top management is unaware of what DW is and the advantages it brings. The DW must also be able to supply the required statistics to government and in-house information for strategic planning and decision making.

This paper reports on how one institution in South Africa, The University of the Free State, tackled and overcame the challenges. First some background will be provided on the history of the current MIS systems. That will be followed by a detailed discussion of the challenges involved to get a DW up and running and how the challenges were overcame. The paper then concludes with the lessons learned which can be applied by other universities in a similar position.

2 Background

2.1 Old IBM system

In 1986 an in-house system was developed by using the IBM platform to fulfil the requirements for data storage and retrieval. With the passage of time the system became inadequate to accomplish its tasks. Some of the major reasons for system failure were programming languages that became obsolete, developers having left the University and the system which was designed in patches and phases ended with lack of data integrity and inconsistency in system interface designing.

2.2 OLTP system from an international company

After the bad experience with the IBM in-house development, the University was ardent to buy a system from some international company, because more and more universities were opting to implement integrated software packages from this company. In 2003 the University purchased a new OLTP system at huge cost by considering the following factors: a complete package with design consistency, analytical and strategic reporting capability, new technology and full technical and maintenance support.

During the course of data transfer from the old to the new system (with a different database design) numerous data conversion errors generated anomalies and a lack of integrity in the database. The new system also proved inadequate to provide the necessary statistics.

Within one year after the installation of the new OLTP system, the University faced a number of new challenges that they have never considered when purchasing this commercial product. The main problem was the lack of customization of the product and they are now not in a position to afford the customization costs.

2.3 HEMIS analyzer

The operational system proved inadequate to provide the necessary statistics to the Department of Education (DOE), therefore, the Planning Unit of the University purchased a new system, HEMIS analyzer, from a local company at the end of 2005. The HEMIS system is basically designed according to the format specified by the DOE. Data is uploaded into the system from ASCII files which the University generates to provide unit record statistics of students and personnel to the DOE (which uses it to allocate the subsidies to the universities). This system provided a workable solution, but with changes in requirements, new reports must be developed and it was worthless for institutional planning and forecasting purposes.

3 Challenges

The following sections will provide details on the challenges that were faced during the development of a Student Data Mart (SDM).

3.1 Dirty data

The University of the Free State is among one of the oldest universities in SA. It has computerized student records from 1946 up to today. The University was interested to preserve this history data because historical data are necessary for business trend analysis [10]. A decision was made to migrate from the old IBM data to the new OLTP system. During this migration numerous data conversion errors generated anomalies and a lack of data integrity.

3.1.1 Missing academic program and plans

There are numerous combinations of academic programs and academic plans for which students were enrolled in the past. These combinations are no longer valid or do not exist anymore. Therefore one can not trace the degree for which the student was enrolled.

3.1.2 No uniqueness

In the Academic Plan table several academic plans were entered several times with different effective dates. One can even enter a new row with the same date and with the same entries, because no primary key constraint was enforced on any table.

3.1.3 Inconsistency in data

Student demographic information is very important for certain types of analysis like drilling down to the country, state/province and to city level. In the OLTP system there is no way for standardizing cities or other such values. For example the city name Bloemfontein was entered 16 times with different spelling.

3.1.4 Spaces in mandatory columns

It was explained in previous sections that the system did not enforce integrity constraints, but Not Null constraints can be found in most of the tables. While populating history data from the old IBM system, spaces were added in such mandatory columns where no corresponding value was found.

3.1.5 Missing links

When students successfully completed their degrees, single entries were made per plan in the Academic Degree Plan and Academic Degree tables for graduation records. In several cases no corresponding entry was found in the Student Enroll table with the plan on which students received their graduations.

The reason for these missing links was the fact that students enrolled in one academic plan, for example a four year bachelors degree, but after completing certain modules the student wanted to scale down to a three year bachelors degree. The department and administration accepted these requests and awarded a three year degree with a different plan. In the OLTP system there is no way to link these changes where the student was previously enrolled in plan A and now received his degree in plan B. In the same way there are students whose record exists in the Student Enroll table, but no related record exists in the student demographic table.

3.1.6 Wrong entries

The entries that the systems itself identify by picking different academic programs, plans and modules from the front-end were wrong. For example, the academic career for the same module was entered with different academic careers in different tables.

3.1.7 Year and semester modules conflicts

In the University there is a difference between the module start, end or census dates for semester and year modules. The OLTP database structure allows entry for semester modules only. For example, 2061 and 2062 represents the first and the second semester of the year 2006 respectively. This scheme works for entering semester modules' start, end, and census dates, but fails when entering year modules.

Administrative staff endeavours to fix or resolve the above data errors but fail to do so, because it is too difficult to fix the errors in the OLTP system now, unless by truncating data from all of the tables and enforcing integrity constraints for future data. This method is still not cheap, because the University has to get consultancy from the international company again and a huge budget is required to do so.

3.2 Limited budget

In the Introduction it was emphasised that the University works with a limited budget. After several wrong and costly decisions they are wary of any new IT expenditures. Ways and means had to be found to develop a DW with minimum cost.

Wierschem *et al.* [2] pointed out that the development cost of a DW can be reduced by reducing the overall scope of the project and the project can also be broken down into smaller components and developed over a longer period of time. According to a survey by Wagner *et al.* [4] it was concluded that enterprise solutions are not suitable for developing countries. For these reasons it was decided to develop only the Student Data Mart for a start. Up till now not much work has been done on fitting student record data to the dimensional model star schema [11].

To further cut down on costs no special hardware or software was purchased, but existing hardware and software were utilized. Preference should be given to in-house existing staff instead of seeking outside help [12]. This advice was followed and a single person trained in DW was placed in the Planning Unit (MIS Department) of the University in order to develop the SDM. Detail requirements were gathered before and during the development process and therefore there was no chance for misunderstanding in the requirement gathering phase.

3.3 Prototypes of student data mart

In this section the star model for the SDM that is more suitable for extracting data for DOE reports and institutional strategic reporting, forecasting, and

predictive modelling will be discussed. Details are also provided for the extraction, transformation and loading process into the DW. A number of procedures were written to fix the errors and enrich the data with information that is required by the DOE and institutional internal needs, but which is not available in the OLTP system.

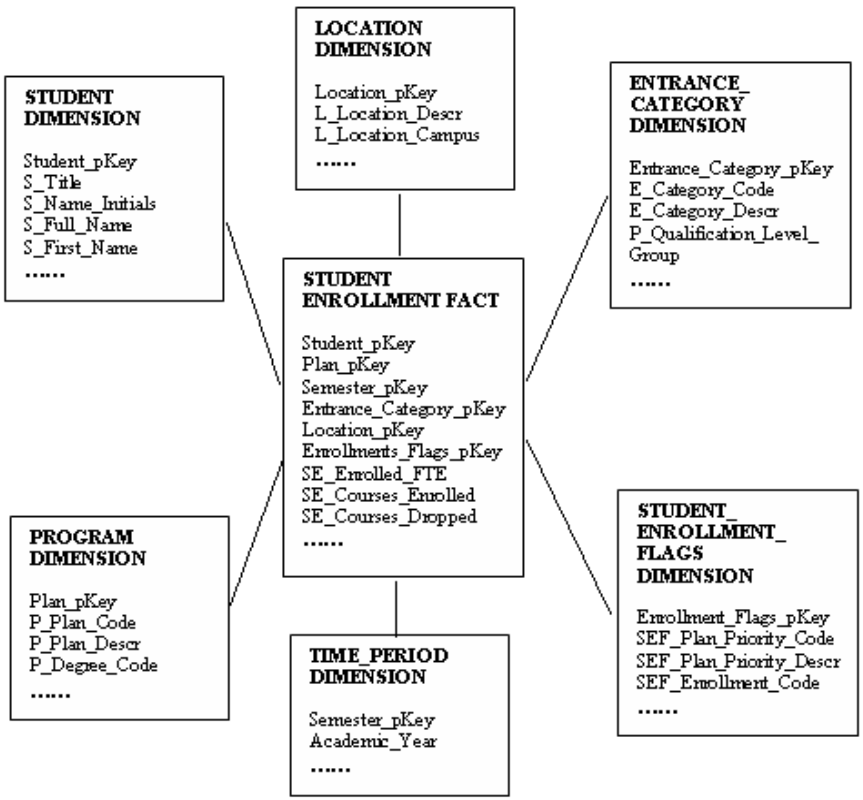

Figure 1: Student enrollment star.

3.3.1 Grain and summarization of the SDW

According to a study [11], the grain of a student record star schema with dimensions of time (academic year, term), student demographics (one record per student) term (one record per student per term) and student matriculation (one record per student per course of study undertaken) would be "student per term per course of study."

To store student enrollments with plans and courses two different fact tables were designed in the current SDM: Student_Enrollment_Fact (see fig. 1) with granularity "student per semester per plan" and Student_Course_Fact (not shown) with granularity "student per semester per plan per course". The main

reason for designing two separate fact tables is the grain and summarization of the fact tables. The Student_Enrollment_Fact store summarized data and it, therefore, contains only 13 percent of the number of rows of the Student_Course_Fact. During the requirement gathering phase it was identified that most of the reports used for reporting and planning purposes were based on the head counts of students in a semester or a year. Student_Enrollment_Fact can answer these queries more rapidly than the Student_Course_Fact.

3.3.2 Student dimension

The student dimension is the biggest dimension in this SDW holding 59 columns. There are fifteen normalized OLTP tables which hold student records. To extract data from these fifteen tables a complex query was required with a number of outer joins and the need to find the last effective date. With this dimension one can drill down and roll up among different hierarchies. Another set of procedures were written to calculate the student's age groups by fixing invalid student's birthdates, native language, nationality, ethnicity and student's resident city.

3.3.3 Junk dimensions

A junk dimension is a convenient grouping of flags and indicators. It is helpful, but not absolutely required [13]. In this project it was concluded that junk dimensions are very useful in the SDM to enrich data for providing certain statistics which are not available in the source OLTP system (see the Student_Enrollment_Flags in fig. 1. The DOE needs student's enrollments together with the primary plan and this information is not available in the source data. A set of filters were used to find students' records which qualify for government subsidies. Some of these filters were: enrolments before the census date, students who had dropped out or withdrawn their plans before the census date, undergraduate students who fulfil their matriculations and students who failed or received re-assessments. This information is processed in the staging area and corresponding flags and indicators are added with the extracted data. A surrogate key is generated in the Student_Enrollment_Flags by extracting unique combinations of flags and indicators. The size of this junk dimension is 0.035 percent of the size of the Student_Enrollment_Fact rows.

From fig. 2 one can see the benefits of using the Student_Enrollment_Flags dimension. Before the SDM it was very difficult to determine the students' primary plans and students who qualified for government subsidies. Now there is no need to write complex queries for finding these statistics, because the Student_Enrollment_Flags dimension makes this task very easy. One can just drag and drop data elements onto the pivot table and get the required results.

4 Conclusion

Current situations in South Africa are forcing HE institutions to enter the DW arena. They face several difficult challenges like low budgets, dirty OLTP data and the requirement to provide statistics to the government and at the same time

have information available for strategic planning and decision making. It has been shown in this paper that the challenges can be overcome by starting small, using in-house knowledge and starting with existing hardware and software. By designing the data marts properly and loading them with clean data, all the required information is available in an easy-to-use manner.

Count of					Acader			
Plan	Enroll	HEMIS	Locatic	Enti	2003	2004	2005	2006
Primary Plan	Enrolled	Subsidy student	Distance	FU	91	257	210	88
			Distance Total		91	257	210	88
			Main	FU	2,840	3,155	3,310	3,319
			Main Total		2,840	3,155	3,310	3,319
			QwaQwa	FU	267	327	454	474
			QwaQwa Total		267	327	454	474
			Vista	FU	220	59	14	13
			Vista Total		220	59	14	13
		Subsidy student Total			3,418	3,798	3,988	3,894
	Enrolled Total				3,418	3,798	3,988	3,894
Primary Plan Total					3,418	3,798	3,988	3,894

Figure 2: Pivot table from student enrollment star.

References

[1] WMU Strategic Plan for Information Technology: Data warehouse proposal, http://www.wmich.edu/sis/pddv1.0.pdf

[2] Wierschem, D., McMillen, J., & McBroom R., What Academia Can Gain From Building A Data Warehouse. EDUCAUSE QUARTERLY, Number 1, 2003.

[3] Hammond, M., Research finds data warehousing market grew 34% in 97 PC Week Online, 1998.

[4] Wagner, C., et al, Enhancing E-Government in Developing Countries: Managing Knowledge through Virtual Communities. The Electronic Journal on Information Systems in Developing Countries, EJISDC, 14,4, pp. 1-20, 2003

[5] Griffiths, S., Data warehousing – what, where, why and how. Data warehousing conference, Johannesburg, South Africa, 12-13 June 1995,

[6] The Use And Value Of Data Warehousing In Higher Education, http://www.mountainplains.org/articles/mpa15.html

[7] Bunting, I., "Funding" in transformation in higher education: Global pressures and local realities in South Africa. Centre for higher education transformation in South Africa (CHET), Pretoria, 2002.

[8] Desruisseaux, P., Universities Venture into Venture Capitalism. The Chronicle of Higher Education, A44, 2000.

[9] Lazerson, M., Wagener, U., & Moneta, L., Like The Cities They Increasingly Resemble, Colleges Must Train and Retain Competent Managers. The Chronicle of Higher Education, A72, 2000.

[10] Data warehousing: Concepts and mechanisms, http://www.svifsi.ch/revue/pages/issues/n991/a991Gatziu.pdf

[11] Data Models For A Registrar's Data Mart, http://www.georgetown.edu/users/allanr/bridge.pdf
[12] 10 rules for successful data warehousing, http://www.kmworld.com/articles/printarticle.aspx?articleid=9081
[13] Kimball Tip#48: De-Clutter with junk (dimensions), http://www.kimballgroup.com/html/designtipsPDF/DesignTips2003/KimballDT48DeClutter.pdf