academicJournals

Vol. 6(1), pp. 1-7, February, 2016
DOI: 10.5897/IJCER2014.0220
Article Number: 43DC51758418
ISSN 2141-6494

http://www.academicjournals.org/IJCER

International Journal of Computer Engineering Research

Full Length Research Paper

Issues in a transactional database emanating from data warehouse project

José A. Z. M.*, Beatriz R. R., Hugo de la R. C. and Patricia R. A.

Institute of Technology of San Luis Potosí, Tecnológico av. s/n Col. UPA, Soledad de Graciano Sanchez San Luis Potosí, México.

Received 6 November, 2014; Accepted 31 December, 2015

When a data warehouse (DW) is built, typical extraction, transformation, and loading operations are made, including cleaning tasks of the data, conflicts resolution, inconsistencies identification, and constrains checking, among others. In this work, two databases (DB) were combined to obtain indexes approved and failing course in basic sciences area (Mathematics, Physics, Chemistry) of the Institute of Technology of San Luis Potosí (ITSLP for their initials in Spanish). The database table intermediate of the transactional database was used to calculate dimensions and fact database tables of one warehouse. Stored data in DW are the input to Data Mining Algorithms and validation information process. This benefit was not expected. At last, the combined database was divided into a transactional database and historical database, to avoid redundancy, inconsistency, incongruity, and other integrity issues and to prepare the historical database to update DW. This DW will be suitable to carry out statistical queries in diverse areas and faculties. The final divisions of mixed database and alternative uses of intermediate tables are some non-awaiting results of original planning of project.

Key words: Data bases, data mining, basic sciences, data warehouse, inconsistency.

INTRODUCTION

The failure courses of students in basic science area (Mathematics, Physics, Chemistry, etc.) are a great concern in most of the institutions of higher education, especially in Institute of Technology of San Luis Potosí (ITSLP), because it represents the peak statistics in a poor school performance. For example, the average of approbatory notes is 59.52% in basic sciences. There is a note of zero if a student does not approve the course versus 30.25% average of the other seven engineering careers in the institute (Computer, Informatics, Mechatronics, Mechanics, Industrial, Electronic, and Electric).

Diverse programs that help students to increase basic sciences area school performance are special courses, teacher-student tutoring and student-student, departmental exams, specialist professor recruiting teaching of mathematics, curricula's normalization, observation and intensive supervision about behavior of

*Corresponding author. E-mail: jazaratemx@yahoo.com.mx.

the classes, and personal monitoring in the classroom to observe the teaching of the titular teachers of the subjects. In spite of all these efforts in the last years, bad notes average was 55.54% in basic sciences and 42.77% in other areas. It did not exist in the ITSLP until this research work focused on the observation and experimentation of statistical patterns that have influence in these mid points.

It is necessary to carry out studies and to elicit data about the possible factors that affect school performance of students in engineering common courses, such as gender, standard of living, social position, and other elements on the transmission of the knowledge process, related with teachers in charge of the courses' attributes.

With the purpose of supporting the analysis of some variables and using an Institute's integrated database (DB), a DW was constructed to apply some data mining techniques, such as correlations, patterns, and diverse variables behaviors that are stored in the transactional databases.

DESCRIPTION OF THE PROBLEM

Some of the educational institutions along the time trust the systems of information like a resource that it supports their processes and they are more efficient when providing access to a high volume of users, students, teaching, and administrative personnel. For that reason, the ITSLP automated the process of storage of data from the students, for the process of inscriptions, queries, and the evolution in their academic progress.

This study revised the current ITSLP database to update the data, which originated several incidences to migrate all information from a database management system (DBMS) to another. In two previous occasions, when integrating the current database with another for their migration with a new one, the essential revision of the integrity, consistency, completeness, and redundancy was omitted as a result of inconsistent, unreliable, voluminous, and incongruous database to manage.

The present study has been carried out with the purpose of integrating, in a correct way, all the information of the historical database with a current database to assure the consistency, reliability, completeness, and non-redundancy, to make it more efficient and more suitable for queries of success or non-success indexes in the diverse careers of the ITSLP.

Another similar DW development projects reviewed by De Witt et al. (2005) and Altareva (2004) explain the benefits derived from the use of DW, but in this case, the project unleashed some issues of the current database and the elder database. These issues changed the way transactional database was considered and they guide us to new challenges.

CONTEXT

Table 1 shows short descriptions of information systems in our institutional lifetime. The development of the systems in Table 1 varied. The first one was developed by a professor of the Department of Computer Systems of the ITSLP, the second one was bought from a former student of the IT of Tabasco, and the third one was developed by the General Directory of Institutes of Technology (DGIT for their initials in Spanish) and distributed for their installation. It was observed that none of the three cases showed a formal effort for the maintenance of each system nor officially presented as an administrator of the database (DBA) in the organizational chart.

During the revision of the databases from legacy and actual ITSLP's information systems (Table 2), classic problems of inconsistency, incompleteness, and no integrity were detected (Zisman and Kramer, 1995). In addition, incongruence in both databases (Pineda, 2015; México, 2015) and database administration of dangerous practices were also found. For example, there is only a single user actual database for many real users and the protocol Authentication, Authorization and Accounting (AAA) has been coded inside the information system instead of using the granted access module of the DBMS.

Until the current system (México, 2015), the integrity checking has been based on the mechanisms provided by the DMBSs, for example, the use of triggers, constraints, primary keys, foreign keys, and domains defined by the user.

Regrettably, these services do not cover the whole definition of database and direct manipulation of the data still remains, without any surveillance of information system or the DBMS. Due to the necessity of guaranteeing better quality possible data for the warehouse, it was especially important to correct these problems in the current database, which requires a great effort in looking for solutions to the problems of the current database:

(1) Inconsistency in the databases overlaps 2006 to 2007 period, because the notes are not the same in the course log table ("the lists") and the transcript of grades table (in Mexico is called "kardex" like the company). This problem appears in both databases and is originated in the bad practice to capture one note in the kardex table, but not updated the list table in cascade to prevent this difference.
(2) Homonymy problems exist among the internal identifiers (ID) of the school subject, because the same ID is used to recognize different school subjects for each database.
(3) Codification issues appear, for example, in a database of 1994 to 2007, the opportunities to take a course are a numeric range from 1 to 7, and in the

Table 1. Evolution of ITSLP´s information systems.

Period	Database management system (DBMS) and information system name	Level of completeness	Data management	Used to meet the needs of these areas
1985-July 1995	Programming Language Pascal	It is ignored that so complete it is.	ASCII-Coded files manipulated by a group of programs	Academic department
August 1995-2007	Clipper X base, Scholar information system SIS	It is missing the data from the period January-June 2001	MS-DOS DBMS based	Academic department and planning office.
2008 until today	DBMS Sybase Integral Information System IIS (Actually in use)	Complete	One of the best DBMS in full edition	Whole institution

current database, they are coded as alphanumerical strings as O1, R1, E1, O2, R2, EE (meaning the first letter type of opportunity and the number denotes how many times the course has been taken).
(4) The concept period is represented as two fields in the database from 1994 to 2006 and in the current is just one; besides that the code varies for the former database in which the period is the year and a digit indicates 1 for August to December semester, 2 for January to June semester, and 3 for the summer period, and in actual database the meaning for 2 and 3 is inverted.
(5) Confusion exists about the career concept that is divided in academic curriculums and these are separated into specialties. This misunderstanding remains in the current database.

DATA WAREHOUSE (DW)

To integrate the information of the second database (Pineda, 2015) with the recent one (México, 2015), classic problems that have been studied in the literature about heterogeneous databases were met (synonymy, homonymy, mismatch type, representation, structure, scale, etc) (Altareva, 2004; Batini et al., 2009; Zisman and Kramer, 1995).

Some solutions to the problems mentioned previously are described as follows:

(1) The problem of inconsistency between the transcript of grades table (kardex) and the notes in the lists solve substituting the notes on the lists for their match in the kardex, in the supposition that the information of this last one is officially considered for the graduation of a student.
(2) The homonymy predicament among the internal keys of the school subject was fixed adding a letter X to them at the end of identical key and upgrading in cascade lists and kardex tables of 2004 to 2007 periods.
(3) The lists and kardex tables were integrated in a single table.
(4) The concept period was represented as two fields in both databases and the meaning of the digit that identifies the semester was unified.
(5) The career concept was revised, and the tables of careers and academic curriculums were normalized and later, an upgrade in cascade in the related tables was made.

Details of the DW design

(1) The star model (Inmon et al., 1998; Kimball, 1998; Jarke et al., 2000) was chosen, because dimensions tables represent more interesting attributes of our final users and the fact table contains principal indicators from these attributes.
(2) The fact table, named *Hechos_grupos*, is based on the information at level course. Its granularity is a course to semester per year; besides, giving us the total succeeded and failed course values, containing other statistics, such as how many of these values correspond to women or man, students with an average less than 70, 80, and 90 at the period to not success, and the same values for the succeeded courses. Finally, this fact table stores the registered students count in the course.
(3) Dimension *Dim_retic* contains columns related with the school subject that was imparted in the

Table 2. Transactional databases from de ITSLP's information systems.

Concept	Database from IIS (2008-2015)	Database from SIS (1995-2007)
Tables	238	57
Indexes	337	63
Stored procedures	628	0
Foreign keys	120	0
Explicit related tables	84	0
User defined domains	34	0
Constraints	120	0
Columns	2837	758
Rows	9402603	785392
Largest table*	Crt_Didactic instrumentation (6,332,368 rows)	DCAR (552,342 rows)
Posible foreign keys	1637	74
Non-explicit related tables	154	10
Incorrect rows in transcript of grades table	79 (almost 0.001% of the total)	12 (almost 0.000025% the total)
Useless rows to enrolment process	87776 (37.37%)	It is not in use
Useless rows to transcript of Grades table	279258 (34.89%)	It is not in use
Useless rows to control courses process	10796 (92.36%)	It is not in use
Empty tables	71	None

course, described in the table of facts, and some of them are its name, number of credits, theoretical and practical hours, name of the career to which it belongs, academic curriculum, correspondent department and suggested semester to take the course.
(4) Dimension *Dim_prof* represents the attributes that describe teachers that were responsible for the course of fact table in a certain period. Fields that compose this entity are the internal ID, the assigned department, category, hired hours, status, gander, grade, master degree and PhD name, marital status, and date of birth.
(5) Dimension *Dim_per* totalizes the calculated data related with each period used to obtain percentage references of data of a group or a set of groups, totalized for one period for academic department, being very useful to compare relative values obtained from the table of facts.
(6) Dimension *Dim_especialidad* is connected with the *Hechos_grupos* and with the dimension *Dim_retic*. Besides, it describes in detail the characteristics of the specialty of a career, as its name, starting date, termination date, maximum and minimum credit quantity by semester, and so on (Figure 1).

NON-EXPECTED PRODUCTS FROM A DW PROJECT

Results described here were derived secondarily from implementation project of a DW. The first ones are related with intermediate tables that were utilized either to obtain statistical values out of the defined in warehouse, or to define alternative dimensions or fact tables of future data marts. The second result was the redesign of the current database, including its migration to another full edition DBMS and the last one was the project of a historic information system, covering 1970, the year of foundation of our institute, until the last semester. The second and third projects are in process, because they exist in a dependency between them due the possibility to reuse the improved design of the actual database in the design of the database of the historic information system.

Useful intermediate tables

It is necessary to generate intermediate tables with data of higher granularity with information that is more detailed and through group of operations to generate the attributes of the *hechos_grupos* table. Alternative utility of these tables is described in the following:

(1) The data of students were integrated from the two available databases (México, 2015; Pineda, 2015) in a new table named *alumnos_nuevo;* one issue exists because the IDs of the students have a mismatch type in the database from 1995 to 1998. This ID is a long integer and in 1998, graduate students are incorporated. Their IDs include a "M" (for master degree), so it was necessary to change the type of ID to char. Querying the *alumnos_nuevo* table can be obtained by calculating values not required in the DW, but useful in the original purpose of analyzing the school performance, such as generation, specialty, semester, origin school, birth

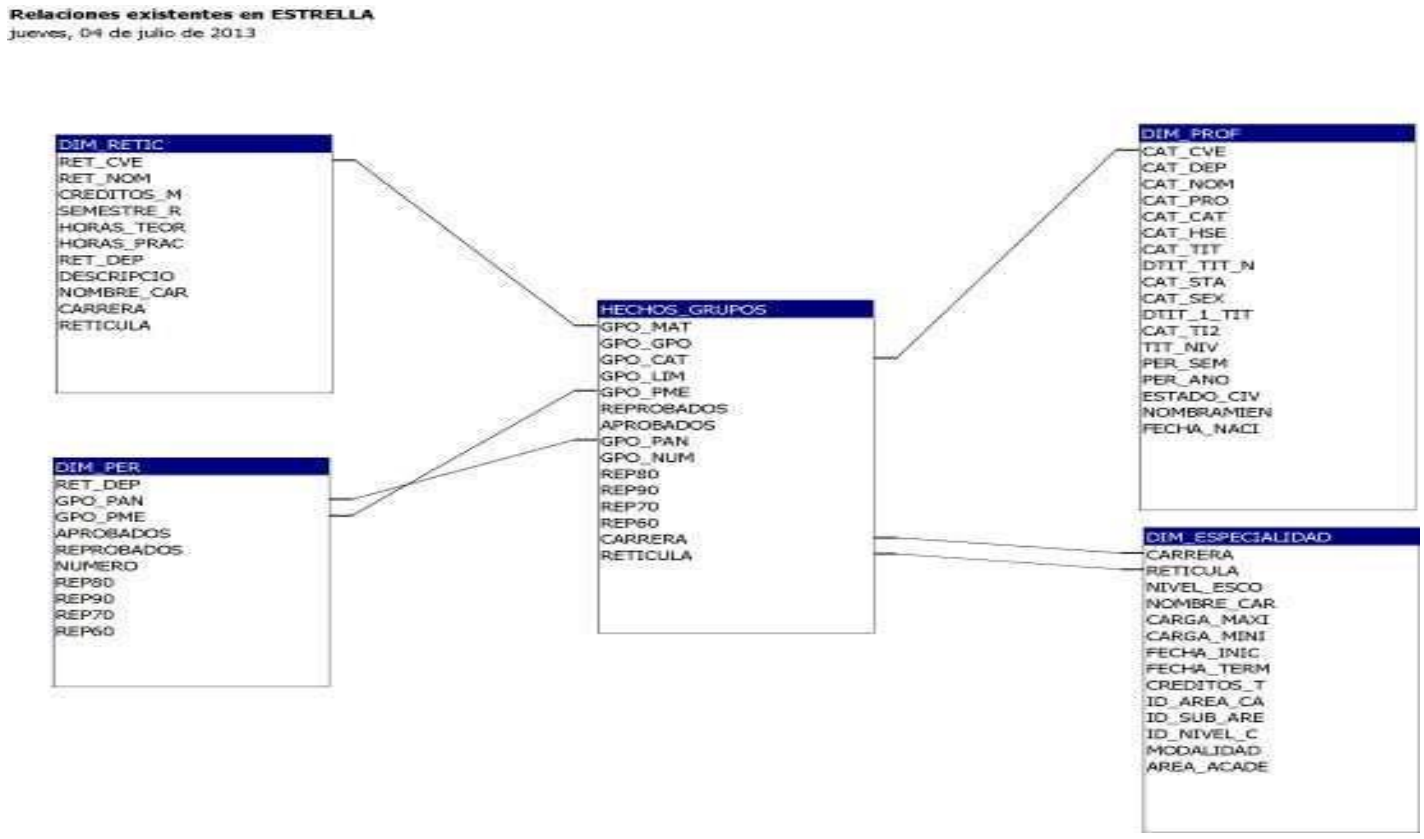

Figure 1. Data Warehouse for the analysis of factors that can lead to non-approval courses.

location, age, marital status, among others.

(2) Another integrated table was *Profesor_nuevo* obtained from joining teacher tables from 1995 to 2007 with the employees table from 2007 to 2012. Employees table was fragmented, because it contains rows of administrative employees and information was recuperated from retired and died teachers. Temporal teachers were a special case, because their information was not always registered in the database and sometimes there was duplicate information of the same teacher in both databases. A serious problem was the different criteria to define internal teacher ID in both databases (Table 2). In the previous database (Pineda, 2015) the primary key was the internal card number and in the current one, it is the Federal Taxpaying Registration (RFC for their initials in Spanish). The form of fixing was to use the card number for all, since in the case of the first database, many RFCs were null. Alternate utility of this table consists in analyzing the professors for attributes as estate of origin, graduate from school, category, assigned department and links with their student's data across the time, etc.

These intermediate tables could be useful to define another fact files in new data marts of the DW used to obtain independent statistical data.

Restructuring and auditory of current database

During the revision of the database of the Integral Information System (IIS for their initials in Spanish) (México, 2015), a part of the DW project had to generate its technical documentation from the *Information_schema* views. It helped us to identify some problems, like those mentioned above, and also facilitated us to discover several irregularities like huge tables. For example, cardinality of largest table (part of the quality assurance system) is more than 6 million rows in just four years in operation.

Some of the main inconveniences in the IIS have been slowness and system crash due to many on-line concurrent transactions. One proposal was a vertical and horizontal fragmentation of some table storing historical and operational information to obtain a transactional database and other historical database, in a similar procedure used to manage historical archive. Some criterion used in selecting and projecting the tables are to separate students' data from graduate and dropout's

data. More conditions were defined but they are subsequently explained in the paper.

We only conserved the course planning, and didactic instrumentation of actual semester and previous one. The others were discarded (the largest table).

To solve the problem of the duplicity of information in the lists table and in the kardex table, the lists kept the actual information of the course and the other stored operational information of the students. When the period ended the data from the lists were passed to the kardex table, and the lists were reset.

Other problems that were outlined with the audit of the current database have a relationship with its maintenance; for example, to drop empty tables, to eliminate repeated indexes and temporary or no longer used tables, to delete incongruous rows (note without course, empty rows, courses without a teacher or without students), in the description of the database of the IIS (Table 2). Some other issues were also described.

The biggest challenge is the inconsistency of data, since you can suppose that the trusty data from 2006 to 1995 come from the SIS system (Pineda, 2015), while those of 2007 until today are kept in current database. But even if this is correct, in the current system it is necessary to revise certain inconsistencies between the notes in the lists of qualifications and kardex (Transcript of Grades table), that occur sometimes with certain courses. For the industrial trainig (although it is not course), the corresponding note is registered in kardex but not in the lists. We repeat the solution used in similar problem presented in the DW construction.

Fragmentation of the actual database will not be enough to avoid the concurrency problems; we propose another kind of fragmentation taking advantage of the natural division between the two areas covered by institute's career: Administrative sciences and engineering. In spite of the teachers involving in these areas, curricula, students, and courses are mutual exclusive. This separation allows managing of 40% alumni in one fragment, and the rest in another.

Obviously, the IIS will be conditioned (México, 2015) for these changes, since at the moment this system should select the data continually, because it has unnecessarily stored a large operational and historical database. Sybase manages the current database in a free-trial version, so that we can export the database to PostgreSQL. To take advance from the skills and experiences obtained from the last two data migrations, we propose to define the database in SQL standard, including statements from Data Definition Language (DDL) and data manipulation language (DML) to avoid work in future changes. In these migrations, we had translated the same data from Pascal records to Xbase rows, later to Sybase rows, and now we have started to change them to PostgreSQL rows (this DBMS make its backup in a flat text file a more suitable way to migrate data).

HISTORIC INFORMATION SYSTEM

It is projected to design the historical database using the new design of the actual database. This historical database will integrate the three available databases. This would imply eliciting all digital information from the academic departments, and is specially complicated in the case of the first information system from 1985, because there is not much experience to recover information stored in Pascal records. Some records from elder database are in flat text, but they have no regular structure, so we start to implement ad hoc programs to pass these records to instructions in SQL, and insert these records in their corresponding tables of the historical database. It is quite necessary to capture the oldest information from the legacy archive kept in paper.

Initial works to implement this historic database are in process but we must redesign the current database. Because the final users have a lot of familiarity with the IIS interfaces, many screens, reports and queries will be reused in this new information system. Potential users are almost the same with current system user's, as academic department, planning office and so on. It also takes advantage of this database to increase the information of actual DW, or the base of new data marts used by other type of users such as education science researchers.

Some strategies to update this database are as follows:

(1) All graduates and dropouts data will be passed to Transcript of Grades table (kardex), together with other records as their personal information.
(2) Do not keep the course planning and didactic instrumentation of any semester.
(3) When the semester ends, information about the curses stored in the list table will be passed from the current database to historical database.
(4) The information of the professors into retirement, died, fired or assigned to another institute will pass to teachers table.
(5) Records liquidated careers, specialties and academic curricula will be stored too.
(6) Another data that only appear in the current system such as traveling expenses, commissions and purchases will be inserted until tax year ends.

Each one of these guides is designed to convert these data in a useful input to update the DW. This database is one of the non-expected results of the original project.

CONCLUSIONS

The current and previous databases were designed by a

qualified computer engineer, but they have been degraded over time due to bad maintenance, urgencies and the lack of a support group that will be responsible for it. Principal contribution of this work is the initiative to redesign conceptual schema and administration of IIS database (México, 2015) derived from DW.

Database technical documentation of the IIS was not found, but a basic one was obtained from Information Schema views, although it must be regenerate after the debugging of actual database to have a definitive version. Technical documentation of system SIS (Pineda, 2015) was found, but it was very limited because it only covers basic information of the tables, such as attribute of column (name, type, length, and primary and foreign keys definition). A lot of obsolete operational documentation of SIS exists, and is only useful to understand some aspects of the older database.

To choose a new DBMS to migrate the current database to it, a benchmark should be made to take the best decision. It could be carried out by teachers and students from the academic computer systems department. It is important to emphasize that these studies of the art should be made continually in any high education institution with a computer systems faculty, because they are very useful for the institution and its students.

A deepest analysis of the design of the current database is in process. In early stages have been detected normalization problem, inconsistency and redundancy. After finishing the technical documentation and the entity relationship diagrams, we can start to redesign the current database, something complicated because 64% of the foreign and primary keys are not explicitly defined (it suppose that percentage will lower considerably after to eliminate useless tables).

The changes that we need to make in current database involve drastic changes in IIS´s queries, not in user interface, as simplification of the access to data must be reached. Both a complete documentation about the database and the information system should be finished. Slowness of IIS and its tendency to crash must to be fixed when this derivate project will be completed. Other details such as a total lack of usability in user interface, a maintenance plan and a complete technical documentation must be corrected.

A reengineering initiative is in progress, that includes a change in the paradigm from PHP to Java, and emigration to another DBMS, in spite of emphasis about a repetitive trouble inherited in the last three information systems, there is the lack of effort in an official project to implement systems covering all development and documentation stages based on a software development model selected.

Conflict of Interests

The authors have not declared any conflict of interests.

REFERENCES

Altareva E (2004). Improving integration quality for heterogeneous data sources (Doctoral dissertation, Heinrich Heine University Düsseldorf).

Batini C, Cappiello C, Francalanci C, Maurino A (2009). Methodologies for data quality assessment and improvement. ACM Comput. Surv. (CSUR) 41(3):16.

De Witt JG, Hampton PM (2005). Development of a data warehouse at an academic health system: Knowing a place for the first time. Acad. Med. 80(11):1019-1025.

Inmon WH, Rudin K, Buss CK, Sousa R (1998). Data warehouse performance. John Wiley & Sons, Inc.

Jarke M, Maurizio L, Yannis V, Panos V (2000). Fundamentals of Data Warehouses. New York: Springer. DOI 10.1007/978-3-662-04138-3.

Kimball R (1998). The data warehouse lifecycle toolkit: expert methods for designing, developing, and deploying data warehouses. John Wiley & Sons.

México (2015). Integral information system. Obtenido de sistema integral de informacón (Spanish): http://www.tecnm.mx/sistema-integral-de-informacion.

Pineda E (2015). Scholar Integration system SIS. from Villahermosa Institute of technology at http://intertec01.itvillahermosa.edu.mx/, Retrieved march 23, 2016.

Zisman A, Kramer J (1995). Towards interoperability in heterogeneous database systems. Available at: https://www.researchgate.net/profile/Jeff_Kramer2/publication/2314437_Towards_Interoperability_in_Heterogeneous_Database_Systems/links/00b7d53c54a5f95acb000000.pdf Retrieved March, 23, 2007.