Building Data Warehouse for the Exam Analysis

Ante Laušić and Boris Vrdoljak
Faculty of Electrical Engineering and Computing
University of Zagreb, Croatia
E-mail: ante_lausic@net.hr; boris.vrdoljak@fer.hr

Abstract: **Data warehousing concept has been developed to satisfy the need in the fastest provision of quality information. The concept was applied mostly in business environments. This paper presents the steps in data warehouse design based on our experience in building the warehouse for monitoring performance of the students on several courses at this Faculty. The need to appraise students' performance at exams emerged after introduction of the new appraisal system. The paper describes the definition of users' requests, the analysis of data sources, multidimensional data warehouse design, the implementation and finally, the usage of the warehouse.**

1. INTRODUCTION

Many different organizations have collected vast amount of data about their business operations and processes. The data are being stored in heterogeneous systems in many different ways, the older ones usually archived to achieve higher performance. It is, therefore, not so easy to quickly obtain all information necessary for taking decisions. The data warehousing concept is designed to tackle this issue.

Data warehousing system is a set of technologies and tools which enable knowledge workers to collect, integrate and flexibly analyze information coming from different sources. The system includes analysis of data sources, data model design, definition of transformation and integration processes, construction of the warehouse and implementation of tools that users employ to get the wanted data from the warehouse.

The paper describes designing of data warehouse that collects data about students' performance on several courses at the Faculty of Electrical Engineering and Computing, University of Zagreb. The subject became increasingly interesting after the Faculty introduced a new system compliant with Bologna Declaration aimed at reforming the structures of higher educational systems in Europe, further development of the comparable degrees system to ensure high standards and improvement of students' mobility within Europe. A course credit system, such as the ECTS (European Credit Transfer System) provides higher flexibility. This Faculty introduced the new system with ECTS on the first academic year with high rate of attendance. A data warehouse for the lectures and exams gives a strong support to the need to obtain a well-balanced system for students' appraisal. The analysis of its data enables end users make decisions about improvement of the system parts,, if they find it necessary.

The following section describes the business processes traced in the warehouse, needs of end users and the analysis of data sources. Section 3 covers multidimensional data model prevalent in data warehousing environment. Section 4 deals with the extraction of data from their sources, data transformation and loading into our data warehouse and different problems emerging during the process. A short review of the ways how to use data warehouse is given at the end of the paper.

2. ANALYSIS

Business process to be modeled must be fully understood before data warehouse is built. End users are interviewed to find out their requirements so that all requirements upon data warehouse are defined. Then data in the source databases are analyzed to see if the requirements of end users are feasible.

Only with full understanding of the business process a designer can distinguish what is important and how the data should be organized in data warehousing system. In our case we traced students' performance on several courses at the Faculty of Electrical Engineering and Computing. For instance, on one of the courses the students can take two mid-term exams during one semester. At the end of the semester all scores, including those for mid-term, lecture attendance and laboratory exercises, are summarized and anyone with the scores exceeding a pre-set limit can take oral exam. The students with enough scores on the oral exam pass the course.

By interviewing end users we learn of their main needs and concerns. At this phase, it should be found out which questions they want to be answered and why it is not possible before the warehouse is designed. Defining end users needs gives the information about the future data warehouse content. The goals of the warehouse should be clearly set, although they might change and be upgraded during design. End users must be contacted at later stages of data warehousing and they must be active in designing the warehouse. End users of our data warehouse are the teaching staff. Their major task is to answer the following questions:

- number of students having passed an exam by the specified date compared to their performance in laboratory exercises and attendance of the lectures,
- percentage of students having passed the exam by taking the mid-term exams,
- percentage of students whose mid-term scores allow them take oral exam, but who decided to take written exam first,
- the grades that students got compared to their success in laboratory exercises and attendance of the lectures,
- comparison of the average grades on different written exams of the same course;
- correlation between success on exams and different high school types the students attended
- correctness of the answers to the questions on the exams; the questions with majority of correct answers and the questions most difficult to answer.

Data warehouse should provide teaching staff with the required information, so that they can analyze their system of lecturing and designing the exams, and recognize and resolve potential problems. By analyzing the answers to different types of questions in the mid-term and final written exams, they can find out where the mistakes usually occur and what lessons require more detailed explanations.

When the requirements of end users are collected, they are checked for their feasibility. Necessary data sources available to the organization in designing data warehouse should be identified. A detailed analysis of source databases must provide the information about available data and their quality. Creation of data warehouse is a process of finding out the balance between users' wishes and soundness of available data.

As a source database we had only one transactional database of the courses and exams. Therefore, we did not have many problems about consistent integration of data into data warehouse.

With the completion of the mentioned tasks and acquired necessary knowledge, it is possible to start designing data warehouse.

3. DESIGN

Data warehouse design is based on a multidimensional data model. The techniques used in the warehouse design are different from those used in designing transactional databases. The survey of data warehouse multidimensional data models that is based on the published papers is given in [1,2].

Golfarelli, Maio and Rizzi present in [3] a conceptual model for data warehousing. In their Dimensional Fact Model, a dimensional scheme consists of a set of fact schemes, and each fact scheme contains a fact, measures, dimensions, and hierarchies. A fact is a focus of interest and its attributes are measures. The dimensions are discrete attributes, which determine the minimum level of granularity chosen to represent the fact. Each dimension is a root of a hierarchy, which determines how the fact may be aggregated and selected significantly for decision-making process. The Dimensional Fact Model is independent of the target logical model.

Multidimensional conceptual data model can be implemented in a relational database or a proprietary structure called multidimensional database. However, end users should never be concerned about the storage of data, and should be able to treat the resulting database as a logically coherent multidimensional structure.

Kimball [4] explains implementation of the multidimensional model in a relational DBMS. The "star schema" is the simplest structure based on the multidimensional data modeling paradigm. It is composed of one table with a multi-part key, called the fact table, and a set of tables with a single-part key, called dimension tables. Each element of the multi-part key in the fact table is in itself a foreign key to a single dimension table. The remaining fields in the fact table are called the "facts". The hierarchy is expressed explicitly in the dimension tables where hierarchical levels are shown as attributes.

In case of multidimensional database, data are stored in the array structure, similar to the programming language array. The array reflects multidimensional conceptual view of data that can be visualized as a cube, if there are exactly three dimensions. Data are segmented into the cells lying in the intersection of dimensions. Dimensions are organized hierarchically, with a possibility to have multiple hierarchies for one dimension. Every dimension acts as an index for identifying the values within the multidimensional array. Physical address of individual cells can be computed using array addressing. Cells in multidimensional structures are often unpopulated, leading to sparse storage.

Relational databases are much more scalable to larger database sizes and it is common that the core of data warehouse is relational and only some smaller data sets needed for On-Line Analytical Processing tools are in proprietary multidimensional databases. Our data warehouse was developed by using Oracle 8.1.6 object-relational database management system. We followed the methodology for designing data warehouse proposed in [4]. In this methodology the most important decisions are made first and decisions are made in the specified order.

First, we had to find out which business processes we wanted to model and what the fact tables were. As already mentioned, in our warehousing project we decided to trace performance on the exams. By interviewing the teaching staff we could understand what information they need. Three different processes were identified and we decided to make three fact tables. Two of them were similar, both traced performance on the exams, one mid-term and the other final (Figure 1).

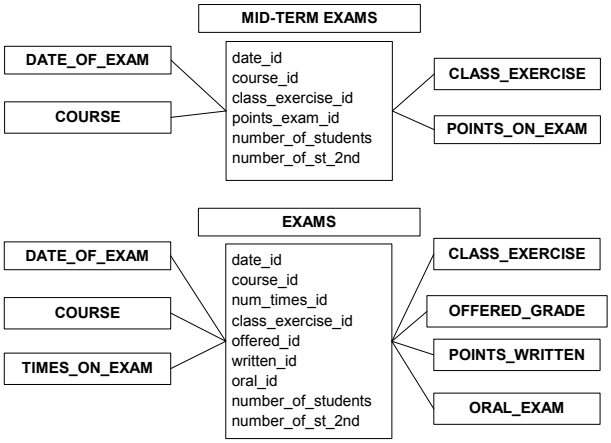

Figure 1. Fact tables MID-TERM EXAMS and EXAMS

After recognizing the processes and fact tables, the grain for each fact table had to be defined. In other words, we defined how detailed the fact table would be. For example, we decided that the grain of the fact table MID-TERM EXAMS would be: "The number of students who got certain number of points on exam at a certain course on the given date according to their success in classes and exercises ". It must be pointed out that we tracked separately the students who attended the course for the first and the second time. After defining the grain of each fact table, dimensions and facts of these fact tables become obvious. For example, dimensions of the fact table MID-TERM EXAMS were POINTS_ON_EXAM, DATE_OF_EXAM, COURSE and CLASS_EXERCISE, and NUMBER_OF_STUDENTS was the fact. For every dimension we had to define as many dimensional attributes as possible.

Figure 2 presents the fact table used to analyze how the students answered different questions and solved the tasks on the exams.

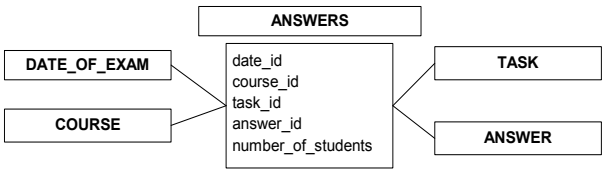

Figure 2. The ANSWER fact table

On every exam there are a number of problems to be solved. The students answer by rounding the answer they think is correct. This fact table gives the information about the number of students who rounded certain answer relating to relevant task on the exam from the relevant course taken on the given date.

4. IMPLEMENTATION

When the logical model is designed and implemented in the database, it should be defined how data from the sources will be loaded into data warehouse. This process is known as extraction, transformation and loading process (ETL). During this process, data from sources are physically and logically transformed. (Figure 3).

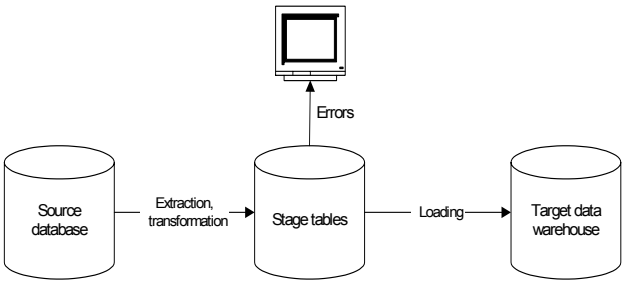

Figure 3. Extraction, transformation and loading

The sources are mostly relational, transaction processing databases or flat files, but also older systems on mainframe computers and external sources. On the other hand, as already mentioned, warehouse data are organized according to multidimensional data model, and stored in a relational database, with some parts in special multidimensional databases.

There can be more then one source and the same data can be stored differently in different sources. It is wrong to underestimate the amount of analyses to be performed in order to understand how to transform and integrate data before loading data warehouse. Data should be stored in data warehouse in a consistent way. It is our task to define the procedures that ensure data consistency and integration, and to define how the data will be extracted from the sources, transformed and then loaded in data warehouse.

There are two modes of loading data in data warehouse. The first one is initial loading, i.e. loading of data in data warehouse for the first time. It is time-consuming and loads large amount of data in both dimensions and fact tables. The second one is incremental periodical loading and refreshing of data in data warehouse. During incremental loading smaller part of data is loaded and the existing aggregate

tables in the warehouse are refreshed. Usually, only the data from the fact tables are loaded. They are then tested, validated and published and will not be updated again. In our case, new data were loaded in the warehouse after each exam and processed.

Extraction, transformation and loading process is managed by scripts in certain languages, such as PL/SQL, SQL, etc. Writing scripts can be a long process, but there are some software products which automate it. These products have graphical interface in which a user specifies the needs and the product generates the corresponding code. We used Oracle Warehouse Builder 3i software package for developing our data warehouse. The example of specifying scripts to define ETL process is shown in Figure 4.

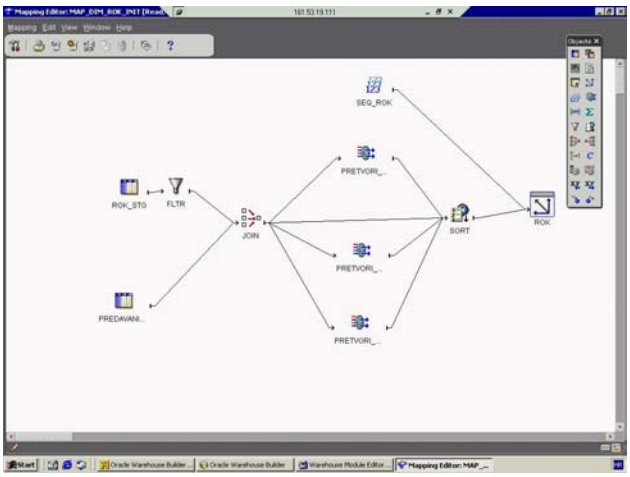

Figure 4. Defining a mapping

These scripts are often called the "mappings". For instance, when we write the scripts that load data in dimensional tables, we map a dimension. Similarly, when we load data in fact tables we map a fact table. The processes of mapping a dimension and mapping a fact table are similar, only the fact table mappings are often more complex.

When we designed the mappings we had to define the joins of tables, transform some data values and make calculations. We also had to prevent invalid values from entering the warehouse and maintain the control over the missing values.

When we defined which source attributes would map to which target attributes, we had to determine which tables from source database we would need to join in order to get proper attributes. We had to do that for every dimension and fact table. For the fact tables we had to determine which attributes would be used in aggregations.

We needed to transform data from some attributes to get the proper result. For example, in source database there was an attribute called TYPE_OF_EXAM. It could contain only two different values: MID_TERM and WRITTEN. But in our data warehouse we had to know if certain exam was a first mid-term, a second mid-term or a written exam, so we had to write a function which transformed data from the source attribute in the proper way.

There were also some attributes in the warehouse, whose data could be obtained only with some kind of calculation of the source attributes. For example, the attribute YEAR in the dimension DATE_OF_EXAM was populated with the years calculated from the date of every exam.

Next problem was to check the data for NULL values, invalid values or missing data, when extracting data from the source. For instance, in the source database there is information about the teacher who designed a specific exam and the teacher who checked it. These two attributes can contain NULL or invalid values and we use these attributes in the warehouse. Therefore, we made a function which tested the incoming data to see if it contained invalid data or of the data were missing.

While designing mappings for adding new data into the warehouse, we had to ensure that only new data were loaded in the warehouse. When we were mapping fact tables, we defined that every mapping had two input parameters: date of exam and course. These two parameters uniquely defined the exam. Therefore, when running those mappings only the data for certain exam were loaded. Mapping of dimensional tables was bigger problem because we had to select all available data from the sources and then to determine which were going to be load. By configuring some parameters, we defined for every mapping which operation would be performed.

Data are extracted, transformed and loaded in the warehouse only after the mappings are deployed to the database and after they are run. Running order of the mappings is important for initial load. All dimension mappings must be run first, and then fact tables mappings. The reason for that is that the fact table references the data in the dimensions, and therefore these data must be available in advance.

5. USAGE

When the data are loaded into the warehouse, the users can write queries, make reports, graphs etc. Different kind of tools can be used to analyze the warehouse data:

- query and report tools,
- On-Line Analytical Processing (OLAP), and
- data mining tools.

In the first phase of designing data warehouse for analyzing students' performance on the exams, we used simple query

and report tools for querying the warehouse, but we planned to use OLAP and data mining tools too. Most users would be satisfied with a set of predefined reports, but the so-called "power users" might prefer many features of OLAP and data mining tools in getting the answers to a wide range of questions.

OLAP is the technology that describes the applications requiring multidimensional analysis of data. End users of OLAP applications [5,6] interactively compose their own queries using simple user interface, similar to a spreadsheet interface. These users should be able to analyze data across any dimension, at any level of aggregation, with equal functionality and ease.

Multidimensional hierarchies of aggregate data enable:

- viewing of subsets of data,
- navigating among levels of data ranging from the most summarized to the most detailed levels,
- viewing data from different perspectives by dynamically changing the dimensional orientation,
- comparisons among different data elements,
- exception reports to highlight unusual situations, and
- time-series analysis to identify trends, etc.

Instead of OLAP user navigating data, the data mining tools automatically discover the interesting patterns. Data mining involves sophisticated algorithms and data analysis techniques.

6. CONCLUSION

Building of data warehouse is a complex and long lasting process, but having a warehouse that can offer high quality information at the right time is of vital importance in improving performance of any organization. In this paper we described basic steps in building a warehouse based on the experience gained during development and implementation of the warehouse that collects data about students' performance at specific courses.

Strict abidance to the convenient methodology is essential for the success of a data warehousing project. After specifying the requirements and analyzing data source, a data model can be made and the extraction, transformation and loading process defined. Upon initial load, data are periodically loaded into the data warehouse and available for use in different ways by end users. Besides simple query and report tools, OLAP tools are very often applied in data warehousing environment, as well as more complicated data mining tools that discover previously unknown data patterns. Building applications for end users' to access the warehouse data is final phase of our project. A portal with reports and OLAP applications for the teaching staff provides simple use of the warehouse because it does not require client installation.

Building of data warehouse is an iterative process where end users play important role in defining the requirements and validating the results. Interaction with users, their training and transfer of experience are prerequisites for frequent usage of data warehouse, and therefore, the success of data warehousing project.

REFERENCES

[1] M. Blascha, C. Sapia, G. Höffling, B. Dinter: "Finding your way through multidimensional data models", In Proc. of ACM 1st Int. Workshop on Data Warehousing and OLAP (DOLAP), Washington D.C., USA, 1998.

[2] Alberto Abelló, José Samos, and Fèlix Saltor: "A Framework for the Classification and Description of Multidimensional Data Models", In Proc. Of 12th Int. Conf. on Database and Expert Systems Applications (DEXA 2001), p. 668-677, Munich (Germany), 2001.

[3] M. Golfarelli, D. Maio, S. Rizzi: "The Dimensional Fact Model: A Conceptal Model for Data Warehouses", Int. Journal of Cooperative Information Systems, 7, p. 215-247, 1998.

[4] R. Kimball. The Data Warehouse Toolkit. John Wiley & Sons, 1996.

[5] E.F. Codd, S.B. Codd, C.T. Salley. Providing OLAP to User-Analysts: An IT Mandate. E.F. Codd & Associates, White Paper, 1993.

[6] B. Dinter, C. Sapia, G. Hoefling, M. Blaschka. The OLAP Market: State of the Art and Research Issues. Proceeding of the ACM First International Workshop on Data Warehousing and OLAP, 1998.