Data warehouse for an e-learning platform

ZORAN NEBIČ¹, VILJAN MAHNIČ²
¹Faculty of Engineering, University of Rijeka
Vukovarska 58, Rijeka, CROATIA
²Faculty of Computer and Information Science, University of Ljubljana
Tržaška 25, Ljubljana, SLOVENIA
znebic@riteh.hr, viljan.mahnic@fri.uni-lj.si

Abstract: - E-learning is becoming more and more prominent at many higher education institutions. While this development has many advantages it also presents the teachers and the administration with new problems and questions, especially related to monitoring and analysis of platform use, including usage statistics, level of adoption at an institution, course activity, among others. We try to answer some of these questions by identifying related indicators and by offering tools that allow analysis and interpretation of results through data warehouse and OLAP technologies, and by building the basis for a future data warehouse that will support decision making. The steps for building a working data warehouse solution and several corresponding analyses are shown using a Moodle e-learning platform used at the University of Rijeka.

Key-Words: - Data warehouse, OLAP, e-learning, Moodle, University of Rijeka

1 Introduction

University of Rijeka is one of the seven universities in Croatia, middle-sized with respect to number of students (≈ 17 500) and academics (≈ 1 100). The University consists of fourteen constituent institutions – faculties, departments, an academy of arts, a library, and a student’s center. It was formally founded in 1973, although the origins of higher education in this area date back to the 17th century.

The University of Rijeka envisions its profile as a research oriented university geared toward sustainable development with high quality and strong performance in outcome-based and lifelong teaching and learning. The educational part of the vision was specifically elaborated through the Strategy for e-learning implementation at the University of Rijeka, and this document became an integral part of the University Strategy.

To increase the efficiency of the university services related to teaching and learning and to improve its overall quality, the specific strategic objectives were set up, among which we can note: to increase the efficiency of studying and to improve quality of teaching. Introducing e-learning could help achieve those goals, but could also help attract more students, reaching new target groups and achieving higher flexibility of the teaching and learning process.

2 Learning management system - MudRi

[1 p. 5] defines a Learning Management System (LMS) as „a broad term that is used for a wide range of systems that organize and provide access to online learning services for students, teachers, and administrators“.

The LMS at the University of Rijeka was put in production in November 2008, and named MudRi, after „Moodle“ and „Rijeka“, which at the same time means „wise“ in Croatian. It is based on the Moodle 1.9.5+ platform, and running on the IBM Quad Core server. Authentication is done through the AAI@EduHr infrastructure: integrated service connecting distributed directory servers of individual educational institutions. AAI@EduHr enables automatic creation of new users at the time of first user login, which makes it easier to add users to the platform, but at the same time allows access to users from institutions that are not members of the University of Rijeka, thus increasing the dependency on external systems, and possibility of introducing erroneous data (e.g. multiple versions of institution names).

3 Reasons behind using data warehouse technology for e-learning

In the recent years there has been a sharp rise in the usage of e-learning systems in the context of higher education as a means to augment the traditional teaching and learning process, or to fully virtualize the course. In the case of University of Rijeka we have seen the number of e-learning courses rising from 0 in November 2008 to 334 in May 2010 in our LMS alone. While this rise is showing progress towards certain goals (e.g. fulfill a strategy benchmark, reach the goal of a project), and possibly is helping with the exponentially rising demand for higher education [2 p. 8], it does not help answer the questions about the teaching aspects of e-
learning. We need tools to help track and measure the activities performed by all learners, to evaluate teacher actions and effectiveness of the structure of the course. At this point we turn to data warehouse technology for help.

As opposed to the database of the LMS, which sole purpose is to support the functioning of the application, the data warehouse needs to have all the information consistent, organized and standardized. Therefore a new data repository following the dimensional model (see section 4.2) is needed to hold the data, after extracting and transformation (see section 5.3).

4 Overview of terminology

4.1 Data warehouse
In this paper we refer to data warehouse as defined by [4]: it is “an analytical database that is used as the foundation of a decision support system...providing intuitive access to information that will be used in making decisions”. On the other hand, we have to consider the term data mart which can be defined as a “subset of the data warehouse...for a particular business function” [4]. In the context of a university, e-learning function would be situated within the education function, which would lead to the logical conclusion that analytical solution built around this function could fall under the meaning of a data mart, but for the purpose of this paper we will refer to the e-learning analytical solution as data warehouse.

4.2 The Multidimensional Model
In order for data warehouses to perform best at analyzing and summarizing data they are based on a multidimensional model. This model views data in an n-dimensional space, usually called a data cube or a hypercube [5 p. 43]. An example of a data cube can be seen in Figure 1.

A data cube is primarily defined by dimensions – various perspectives used to analyze the data, and facts – cells of the table. In the data cube in Figure 1, which is used to analyze user activity in e-learning courses, we show three dimensions: Activity Type, Time and Faculty a user is coming from. Dimensions have members, so Medical School, Faculty of Engineering, Faculty of Philosophy, and Faculty of Teacher Education, are all members of the Faculty dimension. Activity Type dimension separates activities into distinct groups according to tools used: Forum, Message, Course and Resource, while the Time dimension represents the quarters of the year. The cells of the data cube have associated numeric values, called measures. The measures shown in Figure 1 represent data about a certain type of activity, during a time period of a quarter of a year, for each of the faculties. The measures in Figure 1 are shown in thousands.

4.3 Hierarchies
The data from a cube needs to be available in several different levels of detail, which is made possible by the use of hierarchies. In a hierarchy a child is connected to a parent, and parent level contains all the information of lower child levels. In Figure 1 most easily identifiable hierarchy would exist within the Time dimension: quarters could be children of the Year level, and parents to the month level.

In real-world scenarios several kinds of hierarchies can be distinguished: if there is the same number of levels between each individual member to the root of hierarchy then the hierarchy is balanced (as would be with the aforementioned Time dimension). Hierarchies which do not conform to this are called unbalanced. One example of an unbalanced hierarchy can be found within the Course dimension of the MudRi LMS, as depicted in Figure 2.

Figure 2 - Unbalanced hierarchy between levels of organization within the University

Figure 2 shows a portion of the organizational structure of the University of Rijeka and its faculties. Courses (not shown in Figure 2) belong to chairs, which in turn usually belong to departments, but can be directly
4.4 Measure Aggregation
Measures need to be aggregated when changing the level of abstraction, for instance, if we change the view of a cube to the All level of the Faculty dimension we will get data for the whole University. Fast aggregation mechanisms are crucial for proper functioning of OLAP tools, since they need to allow users interactive exploration of a cube at different levels of abstraction.

5 E-Learning Data Warehouse Development Cycle
It is advisable to have a process model which allows us to know the different tasks that have to be done and the deliverables that must be obtained. [3 p. 7] proposes a lifecycle similar to the one shown in Figure 3:

![Figure 3: E-learning Data Warehouse Development Lifecycle](image)

Some authors go into more detail when listing phases of development, while others, for instance [5 p. 256], summarize the list into four distinct phases. However, [3] and most other authors list, in one way or another, the following activities:

1. **Requirements specification** is a phase that is often overlooked, which leads to the situation as mentioned in [5 p. 259]: “it is estimated that more than 80% of data warehouse projects fail to meet user needs and do not deliver the expected support for the decision-making process”. As a result of this phase we can usually produce a document with example questions and reports, providing the necessary elements for building the data warehouse schema.

2. **Dimensional modeling** includes designing a multidimensional schema based on the analysis of user requirements, and transferring it to the underlying system, usually a relational database.

3. **ETL Design and Development** includes designing the data stage schema, defining the Extraction-Transformation-Loading (ETL) processes, and synchronization settings between the source of data and the data warehouse. The ETL process can be the most difficult part of developing a data warehouse [5 p. 294].

4. The main purpose of **OLAP Cubes** phase is to design and build the On-line Analytical Processing (OLAP) cubes needed to provide the data for reports defined in the step 1.

5. **User interface development** phase results in the user interface allowing the end user to create reports, and browse the data from the data warehouse.

6. **Maintenance Plan**

7. **Test and Deployment**.

In this paper we focus on the first four phases of building the system, and look into them more thoroughly in the following sections.

### 5.1 Business Requirements
As noted in section 5 the result of the initial phase of e-learning data warehouse development is the requirements document including the questions that need to be answered. In this paper we focus on several questions from the domain of the University management (platform growth, teacher activity), system maintenance (workload), and teachers (student activity):

- Regarding the course: when are new courses opened? Reporting by institutions and time.
- Regarding the student: when do new users register? Rise in student numbers by institutions and time. Student activity by type, time of day and year. Usage of collaboration tools.
- Regarding the teacher: measuring activity, correlating activity with number of students, and learning outcomes and grades.

We will try to offer answers for some of these questions in section 7.

### 5.2 Designing a multidimensional schema
The designed data warehouse follows the multidimensional model introduced in section 4, in
which a dimensional schema is made up of a central fact table and dimensional tables which help analyze the data.

Questions posed in section 5.1 can be answered using different multidimensional star schemas. One example of a developed star schema is shown in Figure 4, which helps analyze the user activity on the platform.

![FactUserActivity Schema](image)

The fact table contains the data we want to know about the action on the course: what action has been done, from which IP address, error code etc. The grain of the factUserActivity is one recorded activity of a user. The fact table also contains the foreign keys that connect to dimension tables.

The defined dimensions of this star schema are: (a) DateDim: generated date dimension, having usual information about the each day of the year (Day of the year, English name of the Day, English name of the Month, Quarter, etc.), with additional information applicable to our situation (whether a day is a holiday in Croatia, name of the holiday, academic year, semester). Although it is generally recommended to use surrogate keys in dimensions [6 p. 60], it was decided to use natural keys, to ease making connections to dimensions; (b) UserDim: user dimension, holding data about the authenticated user of the platform: means of authentication, username, first and last name, affiliation, day of joining, date of last access, e-mail and other contact information, and a number of other attributes; (c) CourseDim: course dimension, data about the parent category (see section 4.3), full name of the course, visibility, date of creation and numerous others; (d) TimeDim: generated time dimension, with minute as the highest level of granularity.

### 5.3 ETL Process

The development of ETL processes start after the dimensional schema has been established. While it is common for data warehouses to contain information from numerous sources (multiple databases, logging systems, external services, to name a few) for the creation of a data warehouse for our e-learning platform we used only the data from the MudRi LMS platform database.

MudRi is running on open-source platform Moodle, and using the MySQL database management system (DBMS). Moodle as a platform is still burdened with legacy limitations of the MySQL Database Management System (DBMS), which initially did not support foreign key referential integrity. Database schema can therefore be obtained only through reverse engineering, which leaves a number of tables disconnected from its database context. The schema is relatively complex, with its 284 tables, so the first stage of ETL included sorting out the relationships and selecting the needed tables. In the following steps of this process a number of activities were performed: (1) extracting data from the MySQL DBMS by means of a MySQL textual export (MySQL dump); (2) loading the data in Microsoft SQL Server 2008 R2 DBMS through Microsoft SQL Server 2008 Migration Assistant; (3) converting Linux date stamp into Date type in selected tables; (4) handling missing and redundant data; (5) cleaning data, e.g. the multiple affiliation labels for the same institution due to external authentication service, as mentioned in section 2.

Even though the most time and effort was put in the ETL phase of the lifecycle numerous refinements can be introduced in the future, such as: automation of the process, further cleaning erroneous data, etc.

### 5.4 Creating OLAP Cubes

Microsoft SQL Server 2008 Analysis Services was chosen as a software product for developing and deploying OLAP cubes creation for MudRi LMS. The decisive factors in choosing this product were the familiarity of the project team with the Microsoft platform, the acquired license, and the completeness of the suite for handling all the phases of the development lifecycle.
6 Microsoft Analysis Services

Microsoft SQL Server suite includes: Integration Services that supports the ETL processes, Analysis Services providing analytical capabilities by allowing management of OLAP databases, and Reporting Services that is used to define, generate, store, and manage reports. [5 pp. 59-63]

Development of these services is done through a special version of Visual Studio 2008: SQL Server Business Intelligence Development Studio (BIDS). Developing an OLAP cube through BIDS and Analysis Services includes [7]:

1. Creating Data Sources – defining objects containing connection details to a data store
2. Creating Data Source Views – a data source view is a logical data model that exists between the physical source of data and Analysis Services Dimensions and Cubes
3. Creating Dimensions – Dimensions in Analysis Services define the structure of a cube, equivalent to Dimensions elaborated in section 5.2.
4. Creating Cubes – Data Cubes are built according to the multidimensional schema built in the step 2 of the Data Warehouse Development Cycle (see section 5). Each cube must contain at least one fact table mapped to measures (values) in a cube, and one or more dimension tables. One simple Data Cube is shown in Figure 1. After creating and deploying a data cube it is possible to browse the cube by creating test reports as shown in Figure 5.

For the data presentation layer multiple applications can be used, one of which is Excel 2010, where graphs can be created, which is presented in section 7.

7 Producing reports

In section 5.1 we suggested a few questions that could be answered by the use of data warehouse technology, one of which concerned the student activity analyzed from different points of view. In the following sections several analysis are presented, concerning the user activity on MudRi LMS.

7.1 System workload distribution

Without the analysis we can assume that the system is most heavily used in the mornings during weekdays, which can easily be tested by the use of the OLAP cube first presented in Figure 4, and the resulting data presented in Figure 6, with hours on the x axis of the chart, and sum of actions on the y axis.

Figure 6: User activity by hour of day and workday vs. weekend

From the Figure 6 we can see that the system is most heavily used in between 10:00 and 19:00, and it is used more during the week, and less during weekends. Additionally, on weekends students start using it later in the day. We can also deduce that the best time for taking the system offline, e.g. for maintenance, would be in between 2:00 and 4:00 on weekdays, and in between 6:00 and 8:00 on weekends.

7.2 Usage of e-learning tools

Another question we posed in Section 5.1 concerned the usage of e-learning tools. We can again use the existing OLAP cube to find out if we can show existence of certain trends for activity types, for instance we can expect to have more quizzes in the morning, and more private messages exchanged in the afternoon. This data is shown in Figure 7.

Figure 7: Usage of e-learning tools
From Figure 7 we can see that more than 30% of participation in forum-like discussion activity happened between 6:00 and 7:00, and lessons had peaks at certain hours of the day, presumably because of the timetable of an individual course.

7.3 Change in user activity by years

Another analysis that can be done on the data at hand is the trend of change among the types of user activity, which is shown in Figure 8.

By analyzing Figure 8 we could come to conclusion that there was a significant change between the situation in 2008, and subsequent years, but knowing that the system was operational for two months only in 2008, we know that the statistics reflect the limited usage of the time. We can see that years 2009 and 2010 are mostly similar, although in 2010 we do see the rise in the percentage of quiz and forum use, and perhaps a bit less work done on the course alone.

8 Conclusion

Through the process of analyzing the need for a data warehouse solution, going through the steps of the lifecycle of building and deploying the solution, and by generating the reports built upon the data warehouse we have tried to demonstrate complexity and benefits of using a data warehouse for e-learning systems.

Although the system built in this fashion can provide useful information according to previously defined requirements, and dimensional schema built upon those requirements, several areas should be developed for the data warehouse system to become fully functional:

- Adding and redefining user requirements
- Integrating new sources of data: web logs, student records systems, etc.,
- Automation of ETL processes,
- Developing of presentation layer for the end user.

References: