ABSTRACT

Faculty of organization and informatics in Croatia represents an interesting choice for vast number of students. In order to enroll students had to pass the entrance exam that consisted of several groups of questions covering different courses. Based on students’ high school knowledge (and grades) as well as entrance test results (and some additional activities) we used to determine the list of students that were eligible to enroll. Later on (during their study) those students passed some exams as well. This paper describes the solution that was built in order to analyze the success of students from different schools and regions. For that purpose a data warehouse was built that integrated data from several different sources and front-end business intelligence tool was used to build reports that were used to clarify the whole situation.

Categories and Subject Descriptors
H.2.7 [Information Systems]: Database Administration – data warehouse and repository.

General Terms
Design, Experimentation.

Keywords
Data warehouse, Business intelligence, Exams.

1. INTRODUCTION

In the past 15 years (or so) data warehouses have been used extensively to integrate data from different applications in order to enable the data analysis and in order to build reports that contain interesting piece of information. While in the past people used to spend 90% of time preparing the data (they needed) for analysis and only 10% of time analyzing the data, data warehouses have made it possible for number to switch their places. It is clear today that spending 90% of time to build a report is not acceptable (any more) and by using data warehouses producing reports is a matter of seconds.

The answer to the question how this is possible lies in the fact that data warehouses are special type of databases organized in a star (or snowflake) schema that is easy to understand and that can be used to build reports, even for un-experienced users. But to enable this report creation the daunting task called ETL (Extract Transform Load) has to be performed (in the background); ETL usually enhances the data quality and resolves the inconsistencies and other things that make the data (in its original form almost) useless. Once the data is in the data warehouse, OLAP (i.e. front-end) tools can be used to build reports and to analyze the data by using advanced capabilities (drill down, pivot, slice, dice, etc.).

Throughout the history it was a very common scenario that many different applications were built and used within business systems. Once the managers realized what IT can do, without any special planning many applications were built and deployed, but technologies used to build those applications were heterogeneous and incompatible. Although each application was useful in a sense and made it possible for number to switch their places. It is clear today that spending 90% of time to build a report is not acceptable (any more) and by using data warehouses producing reports is a matter of seconds.

SQL is a dominant and standardized language used to work with databases and data warehouses (although some data warehouses can use MDX as well, but we will skip that). Although the language is standardized, during the years it has become quite complex. Further on, when the language was introduced the idea was that it was supposed to be simple so that end-users could build queries on their own; today we know that professionals can have problems with some queries as well and it is not reasonable to expect that end-users pose complex queries that are usually needed to answer some question. Further on, the logical data model is usually complex (containing many tables) and end-users are usually unaware of how to join them together (in different ways) in a single query. Since data can be found in many heterogeneous applications and SQL is complex, user friendly
interface is a “must” in order to build successful reports. Business intelligence tool fit nicely as we will see later on.

This paper focuses on (entrance) exams at the Faculty of organization and informatics, University of Zagreb. The problem that was obvious was that in one point of time we have had five different applications that supported the exams (in one way or another). However, it was almost impossible to build certain (advanced) reports that would show us which students were the best i.e. which schools (regions) produce the best candidates, etc. Because of that a data warehouse has been built and certain reports have helped us to come to some very interesting conclusions.

The rest of the paper is organized as follows; the problem scenario is explained in the following section and then the data warehouse model is explained (briefly). Later on some reports (that were created) are presented and some remarks are given. In the end the conclusion is presented.

2. EXAMS

In order to enroll, students had to pass the entrance exam. Entrance exam contained questions from several different areas (courses) like math, informatics, foreign language, etc. (the list of courses sometimes changed from year to year). The entrance exam was used to check whether students know all the things that should know and based on the entrance exam results some students (i.e. candidates) were eliminated i.e. the best candidates were enrolled.

A very interesting point here is that students (or candidates) that came from certain schools (regions) usually achieved better results meaning that some schools were better than some other schools. Further on, students could get some additional points for certain activities as well (professional athlete, third foreign language, etc.) and it was obvious that certain schools produced candidates with higher number of additional points. Because of that it was crucial to understand what was really going on.

The entrance exam took place two times a year; in the summer (usually in July) and in the autumn (usually in September). After the entrance exam was over we had to build a few reports showing the number of points, average high school grades, etc. Although some of those reports were not complex, they were usually built within a few days (one report required several different queries to be combined and presented on a single report).

If we tried to compare the results to the ones from the previous years, this was not so easy to accomplish. Further on, once the students were enrolled they passed certain exams. Comparing their high school grades with entrance and passed exams results was almost impossible. Five different applications contained relevant data and there was no way to present them in just one report:

- one application was used to store the data on entrance exams;
- one application was used to store the data on high school grades and
- three applications were used to store the data on exams that students had passed.

For the passed exams three different applications were used:

- one old DOS application that was not used but still contained some relevant information on student exams (history data is relevant for data warehouses);
- the successor to that old DOS application was an application that was used for the same purposes for about 4-5 years, and was developed by our local team;
- and the third application (still) used for the same purposes is the one developed and supported on the university level.

In order to analyze the data a small data warehouse was built that integrated data from different sources; the data warehouse model is explained in the following section.

3. THE DATA WAREHOUSE MODEL

As we already know data warehouses are databases that are arranged according to some other design principles; unlike databases where redundancy causes anomalies, in data warehouses redundancy is desirable because it reduces the number of joins and because it makes the model understandable.

The star schema in the data warehouse contains a fact table and dimension tables that are (usually) organized as one can see in the Figure 1:
Since the data warehouse was rather small (one table had 200,000 rows and other tables much less), it was implemented in MS Access 2007 with four different fact tables:

one was used to store data on the high school grades;
one was used to store data on the entrance exams;
one was used to store data on the passed exams, and
one was used to store data on additional points.

We had several dimension tables as well (candidates, courses, additional point, etc) as can be seen in Figure 2 (Business Objects XI was used as a BI tool; the Designer tool was used to build the universe).

One can see (although in Croatian language) many measures and dimension tables with their attributes; all of them can be used to analyze the data and produce reports. This is done in the following section.

More on Business Objects can be found in [1]. More on data warehouses can be found in [2], [3], [4], [5], [6] and [7].

4. REPORTS

We have built a number of reports (what took us hours before was built in a matter of seconds) and some results were astonishing. It is important to have in mind that this warehouse was built within a few months and it has enabled us to build reports that nobody even tried to create before because this was just not feasible (and it was expensive as well). The reason why this was not feasible was that several incompatible technologies were used and nobody tried to extract data from those different systems. With this data warehouse reports that couldn’t be built before have been built within minutes. The first report shows the number of candidates compared by years:

![Figure 3. Number of candidates per year](image)

This report is not complex but shows how the number of candidates changed during the years (the year 2010 is to be observed cautiously because data were not complete). This report is interesting when regions and counties are added; one can easily determine the regions and counties that should be targeted more aggressively to attract more students, especially when we take into account the regions and schools from which the best candidates come.

Another interesting report shows the average number of points achieved per several courses (Hrvatski jezik means Croatian language, Informatika means Informatics and Matematika stands for Math) and month of the entrance exam (7 or 9 i.e. July or September):

![Figure 4. Average number of points per course and month](image)

It is obvious that in July students achieve better results than in September (the maximum number of points is 200) as well as that Math causes the most problems.

Another interesting report is a slight modification of the previous report i.e. it shows the same data only per regions and course (green represents Croatian language, blue represents Math and yellow represents Informatics while the X axis represents the region names in Croatian). Since regions form a hierarchy (region – county – school), the report can be drilled down to county and school, but this will not be shown because it is then obvious which school produces the best candidates (but when this was done the results were surprising):

![Figure 5. Average number of points per region and course](image)

Another report is quite interesting because it shows the number of points achieved through some additional activities (table rows represent 1st – 3rd place in competitions, second foreign language, already passed entrance exam, professional athlete, second high school and third foreign language and columns represent years); each activity brings 20 points. We see that the most influential activity that brings many additional points is the second foreign language. Some cells are empty meaning that (for that specific year) we didn’t have candidates with such type of additional activity:
Another very interesting report shows the high school grades, the entrance exam results and the passed exam results per region. The story repeats again and one can drill down and see the results on the school (county) level, but this is (again) not shown in the paper for obvious reasons:

<table>
<thead>
<tr>
<th>Region</th>
<th>High school (max. 400)</th>
<th>Entrance exam (max. 600)</th>
<th>Passed exams (max. 5.00)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zagreb</td>
<td>209.63</td>
<td>297.73</td>
<td>2.04</td>
</tr>
<tr>
<td>Nepoznato</td>
<td>222.73</td>
<td>262.24</td>
<td>2.07</td>
</tr>
<tr>
<td>Dalmacija</td>
<td>223.21</td>
<td>305.38</td>
<td>2.01</td>
</tr>
<tr>
<td>Središnja</td>
<td>227.08</td>
<td>306.60</td>
<td>2.08</td>
</tr>
<tr>
<td>Gorsko primorska</td>
<td>228.96</td>
<td>341.19</td>
<td>2.28</td>
</tr>
<tr>
<td>Slavonija</td>
<td>231.94</td>
<td>324.23</td>
<td>1.91</td>
</tr>
<tr>
<td>Bosnia</td>
<td>258.57</td>
<td>268.23</td>
<td>2.42</td>
</tr>
</tbody>
</table>

**Figure 6. Additional points**

One can easily say that this data warehouse represents a valuable source of information. Reports in the paper contain important information and could be helpful in many ways i.e. to improve the “input”, to guide marketing activities, etc.

5. CONCLUSION

We can say that the results were (and still are) interesting. Many trends can be spotted and some assumptions that we had had were confirmed.

An interesting thing however is that people that were supposed to be interested in such a solution didn’t show much interest at all. Since they are all familiar with IT, it is hard to explain their behavior.

In our future paper data mining techniques will used since data warehouses represent cleaned collections of data suitable for data mining techniques.

6. REFERENCES


