A Data Mart approach to analysing exam data

Asko Tiidumaa

Department of Information Technology Services; Institute of Computer Science
University of Tartu, Estonia
asko.tiidumaa@ut.ee

Abstract

An organization generates a lot of data by running its everyday business processes. As the organization needs continuous directing and managing based on the previous experience, the constant growth of the data makes the decision making process increasingly more challenging. A solution to this problem is to incorporate data in a cleansed, integrated, analysis-objected manner into a central repository. This paper is about implementing a student exam records data mart as a pilot project of a university-wide central data warehouse.

Keywords: Data warehouse, data mart, data quality

1 Introduction

There are several data sources and information systems available at the University of Tartu. The management and administrative processes often need to view the data as a whole. Also lots of long-running historical reports are run on operational systems. These complex resource-hungry reports often disturb the on-line users of those systems, causing several complaints about poor performance and response times.

With the replacement of the old financial and human resources management system with a new one, some of the reports were implemented using Microsoft Analytical Services. As the flexibility and performance of this multidimensional approach made a good impression on the key persons, an idea of a university-wide multidimensional analysis and reporting system has been arisen.

The current paper presents the results of a small pilot data mart that was used to present the possibilities of the multidimensional analysis to the key persons. One of the objectives of this project was also to evaluate the data warehousing and OLAP capabilities in the Oracle 10g database engine. The paper first describes the project scope and presents the driving questions behind the project. Each of the four sub-objectives will be presented in separate sections of this paper followed by some ideas for future work.

2 Project scope

For the data integration, reporting and analysis purposes a data warehouse has proven to be a good choice [1, 2, 3, 5]. As there was no single estimation on about how large and complicated the university really is, a pilot project has been started. The project has four main objectives:

1. To create a working prototype
2. To gain a better understanding of the technologies involved
3. To get a brief overview of the complexity of the data and the academic legacy
4. To get an overview of the data quality

The working prototype will be used as an example for the upper management and parties involved to show the capabilities of the system in development. Prototyping is also needed for the verification of the design and it allows the end user to get a glimpse of the final product and also suggest some corrections where necessary [6, 7, 10].

The technological aspects should also not be underestimated, as the choice of the software will be a strategic and relatively long-term commitment for the organization. The choice of the data warehousing and analysis software is also crucial because of the shortage of skilled and seasoned partners. The technology that was to be evaluated during this project is Oracle 10g data warehousing and OLAP, since we have Oracle RDBMS as our corporate standard for databases.

The pilot project will concentrate on the exam mark analysis. A top down approach has been chosen, so at first it calls for identifying the driving questions that users tend to ask. These questions have been divided into two groups: quantitative and qualitative questions. There are 11 faculties and four other institutions for which separate statistics need to be maintained. The quantitative questions also need to be answered by the origin and funding schema of the students.

Some of the quantitative questions identified are:

- Number of exams taken
- Number of exam retakes
- Number of exams marked as not attended
- Number of exams failed

- Average number and variance of exams per student
- Credit points given

And the qualitative questions:

- What is the average grade of some subset of the students (i.e. specific faculty, study program, year of enrolment)?
- What is the average of grade of marks given by a specific lecturer and how does it change over time?
- Which subjects and lecturers tend to have a high fail/pass ratio?
- Which subjects or lecturers make the most of the failures for a specific course track?

There should also be a possibility for some ad hoc analysis of data. There should be separate measures for local and foreign students as well as for state-funded and privately funded students.

3 Prototype

By analysing the key questions some measures and dimensions were identified. For the sake of simplicity, the data mart was implemented using a simple star schema. The identified measures or numeric report values are the number of exams and the sum of grade points. In addition, the following dimensions have been identified:

- Student
- Lecturer
- Subject
- Mark/Grade
- Time
- Origin of the student (local or foreign)
- Funding (privately or state funded)

To broaden the possibilities of the ad hoc analysis one can add some extra properties to the dimension data. A common way to enrich a student dimension is to add the information on the secondary school and also some geographic data.

Time dimension can be enriched with the data for the holiday indicator value. Specific to the academic institutions are the indicators of academic year and term – spring or autumn.

Actually, most of the analysis is done in the context of the academic calendar. So it is possible to abandon the classical calendar year hierarchy (levels year/quarter/month) and keep only the simple two-level academic hierarchy of levels academic year and term. Specific to the academic year is that it starts from Monday somewhere near the 1st of September. These kinds of irregular time periods pose a challenge for the analysis tools in performing adequate time-series analysis[1].

To expand the possibilities of the data mart, time dimension and also facts are kept on day level. As the driving questions behind the pilot project require data only on term level, necessary aggregations are done between the data mart and the ROLAP engine.

Grade is considered to be a separate dimension as there is a current need for keeping records for multiple grading schemas simultaneously. The positive and negative grades need also special attention. The grade dimension thereby has two hierarchies: one for grading schema and the other for sign of the grade.

The final number and design of dimensions and cubes related to exam outcome analysis is subject to change in the real implementation of the data mart. The reason behind this is that the OLAP kind of approach seems to gain more benefit from dimensions as from specific dimension attributes. A good example of this is the need for counting exams for students, funded either by the state or privately. This can be done with a dimension attribute for funding on the student dimensions lowest level or as a separate hierarchy. Another option is to make the funding attribute as a separate dimension.

The strict and final definition of the ETL process itself is not an objective of this project. The exam marks come only from one operational data source, the Studies Information System (SIS). As the SIS has currently all the historical data on-line, the minimal ETL process itself consists of several fairly complex SQL statements that query and transform the fact and dimension data on the fly. Building the whole 1,5M exam mark fact table with a full refresh of the dimension values takes a few minutes on the current hardware setup. The time required for doing a full refresh of the MOLAP cube, however, is a completely different story.

4 Technology

The University of Tartu has the Oracle RDBMS as the corporate database standard. Oracle Internet Application Server is also widely used in-house and Oracle Portal is used as a framework for implementing some smaller projects. All of the information systems are currently using one central Oracle 9i R2 database for data storage and retrieval.

This does also mean that the university has some ten years of experience with this vendor's software and communication procedures. Although the evaluation process and pilot project are currently run on the Oracle Business Intelligence (BI) platform, the final choice for the BI tools may change.

---

[1] This has been a minor setback with Oracle OLAP family of tools and is filed as a bug 4323836

The architecture of the data mart is outlined in the Figure 1. The source data originates from the Exam Records module of the Studies Information System, an Oracle 9i R2 Standard Edition database. The Data Mart is implemented on a separate server in an Oracle 10g Enterprise Edition database. The source system is connected via a database link to the data mart and the ETL process consists of several complex SQL statements that transform and cleanse the source data on the fly.

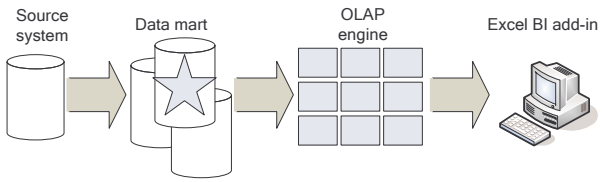

Figure 1: Overview of the data mart architecture

The star schema with its fact and dimension tables can be accessed as a ROLAP cube from the client side. However, to get the reasonable response times for queries against the ROLAP cube, materialized views or pre-aggregated data should be used. One can also write his own SQL statements that often are much faster than the ones from the standard query tools. However, to encourage SQL writing is not an option if we want this system to be available for area analysts and not only to the people with strong technical background.

That's why the MOLAP cube comes into play. The OLAP cube resides also in the Oracle 10g database. The connection between the data mart and the OLAP engine is done via SQL views (stored queries). The mapping of the ROLAP cube into the MOLAP cube is pretty straightforward. However, due to the fact that the data mart has data in the day level and the end users mostly require data on term level, the exam data is aggregated in these SQL views.

One important factor in such analytical applications is also the availability and quality of the metadata. When implementing the cubes in Oracle way, there is a central OLAP Catalogue for the metadata. This makes it easier to access either ROLAP or MOLAP cubes within the single front end tools. However, the previous versions of Oracle Discoverer needed a separate metadata, an End User Layer to access the ROLAP cube. There is also a new version, code named Drake that can also do queries against MOLAP cubes.

The Excel BI add-in does not need the source of metadata separated for those two kinds of OLAP cubes. Why bother, if the driving questions are asked only on term level and this is what the MOLAP cube is built for? We think that the ROLAP cube or the data mart itself can answer the questions in more detail for power users if needed.

The data warehouse by definition is a cleansed, integrated source of data by itself. This makes for a good starting point for conducting some advanced analysis or data mining techniques to get better insight into the process behind the data.

The employees of the university have used to spreadsheets and word processors. The advanced analysis tools often are standalone products that have only a weak connection with wide-spread spreadsheet programs. This link is often done as a data export feature in the analysis tool. Nowadays, several vendors are creating their own OLAP Excel add-ins to preserve the look and feel of the product the end-users love the most [4].

There are several factors that drive the decision, whether to stick with the current software vendor or not. These are price tag, licensing options and the availability of a reliable partner nearby, to name a few. The latter one is the most important criterion because of the shortage of qualified specialists available nearby. As educational as a university is, the learning curve for data warehousing and OLAP analysis tools seems too deep to build up a fully qualified team ourselves.

As we have gone a long way with Oracle database, we feel tempted by implementing all the data warehousing and OLAP systems with solutions from this single vendor. It should be okay to do this with ETL and data warehousing but it is the OLAP engine that has raised some concerns regarding the availability of skilled specialists and documentation. We wrote the words "Oracle" and "OLAP" into the search engine of a major on-line bookstore and received just one reply – a book that is a bit out of date already.

5 Complexity

The academic heritage of the University of Tartu contains some issues that the data warehousing books do not write about:

- Student is allowed to be involved in more than one study program
- A lecturer can be working for more than one faculty
- In some rare cases a study program belongs to two faculties
- The name of a course may vary over time

The environment seems to be different from the academic data warehouses implemented before as the student is allowed to be involved in more than one study program. The exam mark is then recorded regarding each of those study programs. For example, exams in courses of advanced math are taken into account both in statistics and in applied mathematics. There is also a possibility to assign a mark to a specific study program only.

As the current data mart is used only for demonstration purposes, these many-to-many relationships need to be addressed thoroughly in the final product.

6 Data quality

The need for historic overviews is plagued by data quality issues inherited from the previous version of the studies

information system. When implementing the pilot, some specific problems have been discovered:

- Students without any geographical records (addresses)
- Students with more than one actual place of residence
- Exam facts without references to lecturers, students, grade, credit points
- Exam facts with clearly incorrect timestamp (for example, 01/01/0001)
- Course records without names
- Invalid identifier values

Some of those might not seem as issues of data quality at the first sight. If the business rules require a student to have a registered home address then it is a deficiency in data not to have one. The presence of clearly incorrect timestamps is an indicator of other incorrect timestamps recorded that are not so clearly wrong.

One interesting class of errors are invalid identifier values. What makes these interesting is that there is no way other than list boxes in the web-based user interface. A short investigation revealed that these values are changed directly on the database level by some members of the administrative staff to bypass some specific features of the Studies Information System.

Some errors may be more serious than just "wrong data". They reveal the downsides of the complexity of the rules that regulate the education process at the university. Every time a subject will be read a new instance, a subject program for that subject will be created in the system. So every student then subscribes not to the subject directly but for that specific subject program. Also the exam protocols should be associated with the specific subject program for that period. If the exam marks are associated with a wrong subject program, a hard-to-find error will be created.

7 Future work

When starting the implementation project of the university-wide data warehouse, the selection of the skilled partner becomes crucial. This will also influence the final choice for data warehousing and multidimensional analysis platform.

As mentioned earlier, there have been positive results from implementing a larger data warehousing project as a sequence of smaller data marts [3, 5]. This encourages us to follow this initiative and start with a larger analysis of the area of interest. This bigger picture will enable those smaller data marts to be compatible with each other. However, one must take care to ensure that this analysis of the global requirements will not become outdated as those little data marts take months to build.

The prototype and its capabilities described in this paper have created some enthusiasm among the executives and analysts at the university. That interest has revealed some other areas that could benefit from the multidimensional analysis capabilities. These areas, for example, are:

- Student graduation and subject enrolment analysis
- Budget surveillance and publishing
- Lecture room load analysis[2]
- Computer workstation usage analysis

The last area of interest, computer workstation usage analysis, will benefit from the fact that the newly implemented Systems Management Server has a built-in Microsoft SQL Server for its data storage [8]. As its standard reporting capabilities are somewhat limited it will benefit from the drill-down feature of OLAP reporting tools.

As stated earlier in this paper, some of the reports in the new Human Resources system were implemented using the Microsoft Analytical Services. Those shall be implemented or incorporated into the university-wide multidimensional analytical reporting system.

The privacy and security issues have been left unattended so far. This is still an open issue at the University of Tartu as there is no clear definition of the ownership of some specific portions of the data. For example, do the exam marks belong to the university, student or faculty? And if they belong to the faculty then to which of the faculties: the one of the student, the subject or that of the lecturer involved?

However, the strongest challenge seems to be to get the devotion and support from the upper management. At the University of Tartu there have been too many cases where an IS project has become yet another IT project [9].

8 Conclusion

Data Warehousing has been proven to be a good choice for speeding up reporting and integrating data from various sources. The need for such a tool has arisen at the University of Tartu. A pilot project has been started to convince the upper management and the end users to invest their time and money in building a Data Warehouse. The pilot project is based on data mart approach to the exam data from the Studies Information System of the University of Tartu.

The results obtained so far have revealed some serious issues with data quality and the complexity of the academic rules. Solving these issues requires dedicated commitment. The pilot revealed that in spite of the development of the relatively new Studies Information System also aimed on minimising data input errors, the data warehousing project can not be conducted without a data quality audit.

[2] The alleged driving question from the rector: could someone please tell him, how big the university really should be?

Acknowledgements

This work has been supported by the Estonian Science Foundation grant no. 5772.

References

1. Ron Allan. *Data Warehousing and the Student Record Data Mart*. in *EDUCAUSE*. 2000.
2. Walter Weir; Randy Goldenstein, *The University of Nebraska Data Warehouse*.
3. Viljan Mahnic. *Data warehouse: a tool for analysing study process as a value chain*. in *EUNIS 2003*. 2003.
4. Nigel Pendse, *The origins of today's OLAP products*. 2005.
5. Viljan Mahnič; Marko Poženel. *Analyzing enrolment and graduation records using a data warehouse*. in *EUNIS 2004 - IT Innovation in a Changing World*. 2004. Bled, Slovenia.
6. Ralph Kimball; Margy Ross, *The Data Warehouse Toolkit*. Second Edition ed. 2002: John Wiley and Sons.
7. Ralph Kimball; Laura Reeves; Margy Ross; Warren Thornthwaite, *The Data Warehouse Lifecycle Toolkit*. 1998: John Wiley & Sons, Inc.
8. Terje Vellemaa; Aron Jalakas; Asko Tiidumaa. *What do you want to manage today?* in *EUNIS 2005*. 2005.
9. Tiia Lillemaa; Asko Tiidumaa. *Implementation of the Document Management System as an audit for sustainability of procedures within the university*. in *EUNIS 2005*. 2005.
10. Mark D. Scott; David Walls, *7 Steps to Data Warehousing*, in *SQL Server Magazine*. 2000.