COMPARISON OF APPROACHES IN DATA WAREHOUSE DEVELOPMENT IN FINANCIAL SERVICES AND HIGHER EDUCATION

Janis Benefelds, Laila Niedrite
Department of computer science, University of Latvia, Rainja blvd. 19, Riga,Latvia

Keywords: Data warehouse development methodologies, banking, education

Abstract: When a decision to develop a Data Warehouse is made, some sensitive factors should be evaluated to understand the tasks and prioritize them. In this paper we assume that there are common characteristics for companies of similar business activities and different for those with opposite activities. This article looks at the interpretation of the same criteria of two Data Warehouse projects in for-profit (banking) and not-for-profit (higher education) areas. We have used the criteria from (List et al. 2002) to compare the results of the two projects. Each section of the paper describes this set of criteria and development methodology for each of the two areas. An evaluation matrix is provided in Conclusion.

1 INTRODUCTION

More and more Data Warehouse solutions appear in the world all the time. Still, there is a very high percentage of Data Warehouse projects which fail (Frolick & Lindsay 2003). There are a lot of articles providing top ten mistake lists, success factors etc. The existence of material on many 'mistakes to avoid' and lack of comparative research that would help make the best choice between different data warehouse development methodologies served as a starting point for this research. According to Frolick and Lindsay (2003), project failure or success is part of interpretation. The aim of this paper is to compare the meaning of the same criteria and describe methodologies used in Data Warehouse development projects in profit and non-profit companies.

List et al. (2002) give a comparison of three approaches to data warehouse development, based on the methodologies chosen by the authors.. In our two data warehouse projects we have used the user-driven approach for the first project and the process-driven approach for the second one.

To compare Data Warehouse design in different business areas, we used our experience in managing such projects, made an analysis of projects in the same industry and available publications, and interviewed a number of key-people from similar organizations.

The three approaches to the data warehouse development are explained in the section „Related work". The section „Financial Services" presents the main characteristics of Data Warehouse development in "a for-profit company", the section „Higher Education" describes the methodology used in our project and the main characteristics of "a not-for-profit company". The final section "Conclusions" contains a comparison of two case studies and different approaches.

2 RELATED WORK

The data warehouse development methods can be divided into three groups – user-driven, data-driven and process-driven. All three approaches to the data Warehouse development are described and compared by List et al. (2002). Some methods exist in all three approaches, about how the approach can be used in data warehouse development.

For example, the user-driven approach application methodologies are described by Westerman (2001) and (Poe 1996). This approach is also used by Kimball and Ross (2002) in gathering user requirements. The data–driven approach methodologies are described by Kimball and Ross (2002) as their basic method, Golfarelli, Maio and Rizzi (1998), Boehnlein and Ulbrich-vom-Ende (1999), Phipps and Davis (2002) and other authors.

Benefelds J. and Niedrite L. (2004).
COMPARISON OF APPROACHES IN DATA WAREHOUSE DEVELOPMENT IN FINANCIAL SERVICES AND HIGHER EDUCATION.
In *Proceedings of the Sixth International Conference on Enterprise Information Systems*, pages 552-557
DOI: 10.5220/0002618305520557

As far as process-driven approach is concerned, Kimball and Ross (2002) state its importance, Boehnlein and Ulbrich-vom-Ende (2000) describe a method based on the Semantic Object Model technique for process modelling, List et al. (2002) explain their method, which is derived from the „stakeholder driven approach" (Kueng, Wettstein & List 2001).

All approaches have their strengths and weaknesses. The strengths could be described as follows: the *user-driven* approach is the best for finding out the needs of potential users, the *data-driven* approach is the most precisely and formal described and the fastest in getting a data warehouse data model, the *process (or goal)-driven* approach helps to define the key business processes and their characterizing measurements and leads to developing a data model oriented at analyzing these measurements, and thus can be the most suitable for decision support in a particular business area. Among the weaknesses of the approaches we could mention the following: in the case of the user-driven approach the users often do not have a clear understanding of the Data Warehouse and also of the business strategy and the organization's goals. In the data-driven approach the data models, which in some cases are developed in a semi-automated way from the source system data structures, may not contain the necessary data for the analysis (e.g. derived and aggregated data). If the process-driven approach is used, the difficulties arise when the business process measurements are to be defined.

In conclusion, we would like to say that more than one approach has to be used in most cases, to get a data model which would match the analytical needs of an organization. Data warehouse developers have to decide which approach to use as the primary one in a particular project.

3 FINANCIAL SERVICES

A real Data Warehouse project in one of the commercial banks of Latvia is taken as a basis for this paper. This is a universal bank with about half a million of clients, its own branch network and relevant subsidiaries. The financial group deals with all the standard financial products.

The main goal regarding for-profit company is profit. Strategic company's profitability depends on tactical and operational goals (List et al. 2002, Jones & George 2002).

3.1 Development Method and Basic Approach

Our personal experience and related research have proved that goal-driven (or process-driven) method is the right one in for-profit organizations. In comparison with data-driven or user-driven methods, goal-driven method is provided to use in an environment where business processes are designed throughout the company and are combined with business goals (List et al. 2002). Successful for-profit companies usually correspond to that.

For Data Warehouse development the incremental approach (Todman 2001) was evaluated as the most appropriate, because of particular advantages:

- pilot implementation;
- quick wins;
- prioritization;

3.2 Project Sponsor and Organisation

Usage of goal-driven development method together with incremental approach usually leads to a particular type of the project organizational structure, which is very similar to the organization's one and that is nothing new for staff.

Therefore the Data Warehouse project should have a supervisor and adviser, i.e. a project sponsor, from the company's top management. Not having a real project sponsor in the top management is mentioned as one of the 'top mistakes to avoid' in many researches and publications (Kimball & Ross 2002).

As Data Warehouse projects usually cover almost all of the company, it's impossible for one person to keep track of everything. Therefore Data Warehouse project is organized in smaller sub-projects (usually per Data Mart, business process or similar to that). Every sequential incremental sub-project supervises an expert from the business area (owner of tactical goals).

3.3 Application Area

The main Data Warehouse application areas in financial industry are data mining and decision support functionality. Data mining is very common for business directions, which are interesting in analysis of history, forecasting and data correlation issues. In our case the lending and overall clients' history was the goal.

3.4 Timeliness or Maturity

There is no point in having a list of clients or products, if that information is one week old already. Commercial data, especially in finance services, are changing very fast and sometimes very significantly. Some authors describe and some vendors say that they provide 'real-time' Data Warehousing solutions. We have to be very careful with definition of 'real-time' and we would like to use a term 'right-time' instead (Connor 2003).

3.5 End-User Involvement

Any company has top experts where key knowledge, competence and experience can be found. The company's Data Warehouse and its content should be reviewed very carefully under the guidance of these people. The representative of the Data Warehouse project (usually an IT person with good business understanding) taking part in those discussions should be an expert of the same level – otherwise it could lead to the situation, when business experts influence the process, outside their expertise (even, if they believe they are IT experts too). If it's not the same persons, real end-users obligatory should be involved in that project stages, where maturity, end-user application features and related issues are discussed.

3.6 Source and Type of Measures

Especially specific for financial institutions is a very high number of measurements to analyze, majority of them are derivative (like client profitability, average turnovers)So, instead of putting attention on data mapping attention to data transformation, aggregation and interpretation should be paid.

Financial industry's company may have a comparatively high number of source systems because of the wide range of their business activities and because of the interest in the key measurements of any other industry or common knowledge bases. It's all related to business opportunities and threats. It's very common, when one particular or set of some business areas build separate Data Marts

Although there may be a lot of source systems, they usually are similar. Those, which are related to the business processes, are based on the granularity of business object (client, product) and many different dimensions. Those, which are related to the external statistic information, usually are provided by some state institutions or similar for-profit institutions and are pretty similar by the content.

4 HIGHER EDUCATION

The user-driven approach is used in the development of the data warehouse at the University of Latvia. Education belongs to the non-profit sector. The University of Latvia is a higher education establishment with 30,000 students. Its business processes include education, research, finance and the university management - all areas are equally important for the university's successful functioning.

The top management is deeply interested in gaining objective criteria to estimate these and to support making new decisions. The administrative director of the university is the main sponsor for the data warehouse project, also other IT projects in the university are initiated by top management. It is difficult to choose the right priorities for data warehouse development at the university. This is not a profit-oriented business. Therefore, the usual goal for data analysis - the profit is not the case. However, money matters like the study fees and other payments have to be analyzed.

The data sources for the data warehouse are the Student Information System and the Finance System, two Oracle databases, but they are not integrated. The first one is developed in the university; the second is a commercial product. This is typical solution for many universities

In the data warehouse project of the University of Latvia, we applied the user-driven approach. The approach is based on the interviews.

In our methodology we used some ideas from Kimball and Ross (2002), namely, how to organize interviews. These ideas we supplemented with our ideas, how to manage and use the gathered information.

After discussions with the main project sponsor, we defined the groups of interviewees. The potential user groups are the following – the top management, the department leaders and the deans of faculties. The last group is the users whose responsibility is data analysis, the employees from the departments and the administrative staff in the faculties.

The interview content was modified for each group of users.

In the selection of the interview questions we followed the principle – the priority is given to the questions that find out the business objectives and measurements. The questions in the interviews were divided into two groups: „Business goals and influence factors" and „Data analysis demands". The following questions from the interviews could serve as examples:

– What are the goals of your department? What do you want to achieve?

Table1: The fragment of the matrix "Interest scope"–"Interviewees".

	Students	Study programs	Employees	Projects
Rector's assistant	The expected and real number of students by faculties. The number of graduates.	Study programs, their number and development trends	The number of professors, the list of employees by faculties, salaries	The number of projects, financing the projects
Chancellor	The number of students financed by the state and full-paying students	The study fee by study programs	The salaries of the employees, the workload of the staff.	

– What information is the most important in your work?

In each particular interview the questions were chosen from the question list, evaluated, if they are appropriate for the business process analyzed and for the responsible employee.

The answers are summarized in the following matrix: one matrix dimension is „Interest scope"; the second dimension is „Interviewees". The cells of the matrix contain the answers to the questions. The interest scope is the group of similar answers. The developer's responsibility is to define these groups, based on similarity of the answers.

The Table1 represents a fragment of the above mentioned matrix. In the project we defined 19 scopes of interest. This number is too large for making decisions on the Data Warehouse components – Data Mart development priorities.

Therefore, we merged the interest scopes into interest *groups*. For example, we merged „Students" and „PhD students". A new matrix with 11 interest groups was made. The cells of the matrix contain 1, if the answers in the first matrix exist. The Table2 illustrates the new matrix.

This matrix served as a basis for analyzing the number of potential users in each interest group, we also applied coefficients, to emphasize the importance of the needs of a user or user group.

We used following system for the result analysis: **1** - for faculties, **2** - for top management and departments, **1.5** - if the interviewee emphasized the particular issue as of the major priority for them.

As showed the results of information processing after interviews the most popular application areas for potential data warehousing were statistical analysis of different business measurements, the top management also expressed the needs concerning decision support.

The above mentioned interest groups were modelled, using dimensional modelling techniques. Some interest groups from the matrix were modelled as data warehouse dimensions (the light grey columns); some of the others did not have source data (dark grey columns in the matrix). Five of the

Table 2: The matrix "Interest groups"–"Interviewees"

coefficients		Students and PhDs	Employees	The budget planning	The finance resources	Equipments	Study programs	Projects	Administrative documents	Foreign communication	Audit	Other
2	Rector	1	1	1	1	1	1	1	1		1	
2	Rector's assistant	1	1	1		1	1	1	1	1		
2	Chancellor	1	1		1.5	1	1					
	…											
2	The planning department manager	1	1	1				1		1		
	…											
1	Dean of Faculty of Phys. and Math.	1	1		1	1		1				1
1	Dean of Faculty of Pedagogy	1	1	1	1.5							1
	…											
		38	34	9	29.5	24	16	14	8	8	4	13

groups were used to model the data warehouse conceptual model. In this point the information from the first matrix was used.

Based on the previous categorized information, the potential Data Marts were modelled.

We started with one interest group and analyzed in detail all answers in the first matrix, concerning this group. The models include all necessary attributes and facts, to answer the known users' questions. Usually one interest group was implemented with more than one dimensional model.

For the conceptual modelling we used derived ME/R notation (Sapia et al. 1998). The modification concerns the attributes – the attribute group is introduced, to improve the presentation quality of the model.

The data warehouse was planned as a combination of individual data marts with confirmed dimensions. The implementation of the data warehouse was planned as a small subprojects chain, one data mart after other, because the sponsors wanted to gain usable results as fast as possible, so the financing for the next subprojects depends on the subproject's results. Therefore, the next step is the development of the data warehouse architecture bus matrix according to Kimball and Ross (2002).

For the detailed description of the dimensions we used a table, which contains the descriptions of the dimension attributes, the evaluation of corresponding data source attribute data quality.

The table and the bus - matrix are additional information sources for the decision about the development priorities, because they help to estimate the development time. But they are not important for finding out the organization's goals and measurements.

The results of the system analysis were presented to the project sponsors. The suggestions about priorities were made based on number of potential users for particular interest group (not for data mart), the existence of data and their quality, the complexity of data marts and number of common dimensions for the 1st development stage.

The user-driven methodology was applied in the project, and the first data mart developed. Our experience proved the right decision in the early stage of project, which approach to chose.

The weakness of the method, common for all user-driven methods, is the unclear needs. The information necessary for a particular user can dominate the analysis of the organization's needs.

But in the case of the universities the involvement of the users in IT projects and particularly in the data warehouse project is essential, because the potential users are informed and their needs are considered, when the models are developed.

5 SUMMARY AND CONCLUSION

Before starting the comparison we expected very different results because of the different business areas and different approaches.

As we can see from the summary table, our assumption regarding "common characteristics for organizations of similar behavior and different for those with opposite behavior" doesn't come true. Only few qualifications are different for the finance company and the state financed higher education institution

The differences are in the data sources and business measurements, not the data warehouse project organization, e.g. end-user involvement, basic approach, project duration, project sponsor or application area.

We can conclude then, there are no big differences in Data Warehouse project development in organizations with respectively different behavior.

Table 3: The Summary evaluation matrix.

Criteria	Financial services	Higher Education
Development Method	Goal-driven	User driven
Basic Approach	Incremental	Incremental
Project sponsor	Top Management	Top Management
Project organisation	Short sub-projects	Short sub-projects
Application Area	Data mining, decision support	Statistic, decision support
Maturity	Near to Real-time (right-time)	Some data seasonal (study semester), some real-time
End-User involvement	Key experts and users	Key experts and users
Type of Measures	Mostly financial Qualitative (scoring, etc.)	Rating, statistics, financial
Number of Source systems	Many	Few
Type of Source systems	Similar	Different

We are satisfied with the results of our comparison. But still, from our experience, we can say that there are some differences, too. In our next research we will look at how the project external environment (like organization itself, for instance) or similar criteria (Boehm & Turner 2003) affects characteristic of Data Warehouse projects.

REFERENCES

Boehm, B. & Turner, R., 2003. Balancing Agility and Discipline, A Guide for the Perplexed, SERA 2003 Keynote Address Research Review San Francisco, CA. Retrieved October, 2003, from http://acis.lsfk.org/

Support, Prentice Hall, New Jersey.

Boehnlein, M. & Ulbrich-vom-Ende, A., 1999. Deriving Initial Data Warehouse Structures from the Conceptual Data Models of the Underlying Operational Information Systems, in *Proceedings of the ACM Second International Workshop on Data Warehousing and OLAP, DOLAP'1999*, ACM, pp. 15-21.

Boehnlein, M. & Ulbrich-vom-Ende, A., 2000. Business Process Oriented Development of Data Warehouse Structures. In *Proceedings of Data Warehousing 2000*, Physica Verlag.

Connor, M., 2003. A Practical View of Real-Time Warehousing. *Business Intelligence Journal*, Spring 2003, vol. 8, No2. Retrieved September, 2003, from http://www.dw-institute.com/research/display.asp?id=6718

Frolick, M. N. & Lindsay, K., 2003. Critical Factors for Data Warehouse Failure. *Business Intelligence Journal*, Winter 2003, vol. 8, No 1. Retrieved September, 2003, from http://www.dw-institute.com/research/display.asp?id=6592

Golfarelli, M., Maio, D. & Rizzi, S., 1998. Conceptual Design of Data Warehouses from E/R Schemes, in *Proceedings of the 31st Hawaii International Conference on System Sciences HICSS'1998*, IEEE, pp. 334-343.

Jones, G. R. & George, J. M., 2002. *Contemporary Management*, McGraw-Hill/Irwin, 3rd edition.

Kimball, R. & Ross, M., 2002. *The Data Warehouse Toolkit*, John Wiley & Sons, 2nd edition.

Kueng, P., Wettstein, Th. & List, B., 2001. A Holistic Process Performance Analysis through a Process Data Warehouse. *In Proceedings of the American Conference on Information Systems.*

List, B., Bruckner, R. M., Machaczek, K. & Schiefer, J., 2002. A Comparison of Data Warehouse Development Methodologies Case Study of the Process Warehouse. In *DEXA 2002, LNCS 2453*, Springer-Verlag Heidelberg, pp. 203–215.

Phipps, C. & Davis, K. C., 2002. Automating data warehouse conceptual schema design and evaluation, in *Design and Management of Data Warehouses 2002, Proceedings of the 4th Intl. Workshop DMDW'2002*. CEUR Workshop Proceedings, vol. 58, Technical University of Aachen (RWTH), pp. 23-32.

Poe, V., 1996. *Building a Data Warehouse for Decision Support*, Prentice Hall, New Jersey.

Sapia, C., Blaschka, M., Höfling, G. & Dinter, B., 1998. Extending the E/R Model for the Multidimensional Paradigm, in *Proceedings of Advances in Database Technologies, ER '98 Workshops on Data Warehousing and Data Mining, Mobile Data Access, and Collaborative Work Support and Spatio-Temporal Data Management*, LNCS vol. 1552, Springer.

Todman, C., 2001. *Designing a Data Warehouse: Supporting Customer Relationship Management*, Prentice Hall PTR.

Westerman, P., 2001. *Data Warehousing using the Wal-Mart Model*, Morgan Kaufmann.