THESIS OF THE PHD DISSERTATION

DATABASE BASED OPTIMIZATION

in organizing of higher education

by:

KESZTHELYI, András László

Eötvös Loránd University

Supervisor:

Prof. Dr. SEBESTYÉN, György CSc

Budapest

2010

TABLE OF CONTENTS

1. Introduction

1.1. Personal motivation

When I was a university student we used to study and practice computer sciences by Commodore-16 and -64 computers. At those times at our department an IBM-compatible XT stood for high technology. (Main parameters: 4,7 MHz clock speed, 10 MB hard disk, 640 kB RAM). I took part in a research led by the late professor Walter Endrei to develop a database for the history of Hungarian technics. I took part not only in collecting and organizing the data but I also developed the first version of the database as well on the above mentioned XT computer.

At those times we learnt the general basics and algorythms of the dBASE-like file level data management as database management.

Some years later I had attended Bánki Donát Polytech where among others Organizing Information Systems in the field of computer sciences became my main task. That object was an optional one at the final examinations, was taught in 2 lessons per week theory and 2 lessons per week practice. So I began to study both the theory and the practice of database management, particularly data modelling which had been in background. Sorry to say that but it is in background even today.

In spite of living in the so-called informational age, after the integration of our Tech my main object went into the background as well. First it was expelled from the set of the optional objects of the final exams, then the number of its lessons was decreased and it ended up finally. In such a situation I was glad to continue studying the questions of database management in general and the questions of data modelling in particular in a PhD-school under the leadership of professor Sebestyén.

1.2. The importance of the chosen subject

It is a feature of our age of informational societies that not only the amount but the variability of data to be handled digitally have reached a threshold, they are hardly manageable. Parallel to this our dependencies of these data have been growing fast. The forty year old idea and phenomenon of the 'software crisis' became very important for today, we can observe the effects of it in our everyday life.

This is caused by the problem of the large amount only a bit. The main problem is that the tools of the relational database management systems, the data modelling, database planning and developing methods and utilities cannot fit the possibilities and, which is worse, the requisites of the theory.

Everyday experience is that the quality of information systems is poor, including their efficiency and loadability. This is such a serious phenomenon that we have been speaking about software crisis since about the late sixties. This idiom had been first mentioned in 1968, then Dijkstra spoke about it in his Turing Award speech in 1972[1]. The main symptoms of the software crisis are the efficiency, low quality, inadequacy for the requirements.

We can observe these nowadays as well, in spite of that that it seems that every important element and field in database management has a well defined theoretical and practical basis.

We can observe these much more in the higher education. Having the credit system become general, having the number of the students been growing intensively and the requirement to fulfil the international standards result in high level organizational expectations which hardly or never would be sorted out by paper based administration.

1 Dijkstra E. W.: The Humble Programmer. In: Communications of the ACM, 1972., XV. évf. 10. szám, pp. 859-866.

1.3. Antecedents of the subject of the research

Having about a forty year past the theory of database management is well defined and nearly perfect today. It was built on an exact scientific basis by mathematical tools as well. There has been no important theoretical innovation or addition to the original theory for many years. Of course the theory of the database management does have some deficiencies, imperfectness and untapped possibilities, but these cannot hold the responsibility for the phenomenon of the software crisis neither one by one nor together.

Database management has not only the relational way but it had some other approaches as well. It seems that the relational way is the one which is proved by the results of practice.

Relational database management systems have a serious career. They are nearly perfect technically, can be used well. The problem is that there has never existed a relational database management system which would fit all the requirements of the theory. This is not because of scientific reasons and it cannot be discussed in this study. Most members of the family of the relational database management systems with all of their weaknesses are good enough in general circumstances for general purposes.

Project management has grown up and become a separate discipline on the basis of the age-long experiences of organising and controlling of the work of engineering and construction. Especially there have arisen a large number of developing methods, standards, protocols and best practices for the last some decades to help building of information systems. Possible and effective weaknesses of these cannot exclusively be responsible for the weak quality.

All these existing problems can result in poor quality only in unlucky situations combined with serious human errors. They

cannot be and are not exclusively responsible for the low quality.

Originating, creative work is based on perfect modelling not only in case of engineering but in general as well. This is true not only in case of simply supported beam but also in case of the finite element method as well. So it is an example for modelling when a writer gives a description of the society and of the age in his novel (e.g. Derengő hajnal [Looming dawn] by István Fekete) or another one when an artist pictures his inner world (e.g. Szigeti veszedelem [Peril of Szigetvár] by Ignác Konrád). The goodness or perfectness of a model can be considered by comparing it to reality (e.g. Tacoma Narrows Bridge, 1940.)

We do some kind of modelling if we try to give help for a given part of the existing world (e.g. for the administration in higher education) by computer based database. Entities, their attributes and relationships selected into the database are a result of a modelling process. Labelling the elements as 'selected' or 'irrelevant', finding the important correspondences is modelling as well, namely data modelling.

In case of information systems the first main step is building a good conceptual level data model, then the logical and physical level planning based on that model.

Among the above mentioned fields data modelling has the humblest special literature in English and in Hungarian as well. Even the system and phrasing of the basic ideas and categories are not uniform and not generally accepted. This is the main reason of the software quality mentioned above and experienced in everyday life - or more precisely the lack of quality.

Data intensive systems are usually applied in automation of administration. In the course of it the benefits of computers can be used quite well, particularly in case of querying frequently the large amount of stored data by various aspects.

This is very true in higher education where the different administration tasks became more complex. The credit system having been started in which the amount of data to be stored and processed was greatly increased, too.

Here in Hungary two big, important computer systems are used in higher education, the Neptun and the ETR, both of them having a decade long history.

Users of these systems have the general experience that the performance of both of them is much lower than it could be expected, in spite of their decade long history and developement.

2. Research fulfilled

The aim of my research was to find the main causes of the problem I outlined in general and in particular, in order to show a possible solution.

I reviewed the historical background, the theoretical basis and practical realizations of the relational database management, the data modelling and the SSADM method for development, their development and the most important ideas and categories.

In the course of my research I have investigated the role and importance of modelling in general and data modelling in particular and their effect on the quality of information systems.

My goal was to show that a good quality data modelling is necessary (but not enough alone) to develop information systems having high quality and loadability. A data model of good quality is one of the most important factor of the quality of the system to be developed.

Standing on the ground of the theoretical basis and historical antecedents I show the correctness of my train of thought by experiments in an indirect way.

It is not possible to prove a statement in general by reciting a finite number of examples. So the essence of my train of thought is the following: although the results of the measurement made by me are not proofs in an exact scientific meaning of my predicate (they must not be considered to be that) but they may support at least the possibility of my predicate being considered true.

If there is a system in use with a given performance and one can measure a much better performance on a test system working in a much poorer hardware and software environ-

ment than the original one then the cause of the result should be searched for in the difference of the data models.

My example from a field of higher education is a specific one but general enough to support my predicate being true in a wider circle.

Confronting my results with the theory of modelling in general and with data modelling in particular I can state that the role and quality of data modelling determines the quality of information systems first of all. This is not a theorem but a foreknowledge at least.

3. Sources and methods of research

While working on my dissertation I have been taking two main aspects into account. I tried not only to reveal the theoretical basis but to examine the experiences of practice as well. This duality determined both the methods and the sources of the research.

I have revealed the theoretical background on the basis of the Hungarian and English literature.

Important references can be grouped by the following way:

Primary, basic works of main importance with a scientific background. Nearly all of them are in English (e.g. The Relational Model for Database Management[2] by Ted Codd). They are indispensable to understand the beginning and developing and the historical aspects of the discipline.

The first group of works are completed and refined by some other works, mainly papers published in scientific journals or conferences, giving specific completion to one or more details (e.g. Grant's paper on NULL values in ACM SIGMOD[3]). These are essential to understand the fine details of the evolution of the discipline. They are mostly in English.

There are secondary sources which show the present state of the discipline. These can be used even as coursebooks. Maybe in English or in Hungarian, in the latter case either

2 Codd E. F.: The Relational Model for Database Management - version 2. Addison-Wesley Publishing Company, h.n., 1990.

3 Grant J.: Null Values in SQL. In: ACM SIGMOD Record, Association for Computing Machinery Special Interest Group on Management of Data, 2008. szeptember, XXXVII. évf. 9. szám pp. 23-25.

translations or original ones. (e.g. Ullman: Adatbázisrendszerek[4] [A First Course in Database Systems by Ullman & Widom]; Békéssy - Demetrovics: Adatbázis-szerkezetek[5]; books of dr. Raffai.)

Primary, basic scientific works in Hungarian, and only in Hungarian, are the books of dr. Halassy on the theory and practice of three level data modelling. Although data modelling and database planning is discussed by many from many different points of view I cannot find other works except those of dr. Halassy which would discuss this field in a scientific manner.

Data modelling and database planning usually come into focus in connection with CASE-tools which circumstance alone means a problem (too-dependency). Which is more the usage of even the basic ideas and categories is not unambiguous (e.g. the explanation of the conceptual, logical and physical levels). There are absolutely no books which would discuss this field not only in a scientific manner but in its full depth as well.

The mailing lists of the discipline must also be taken into account as the forums of the everyday practitioners. In these sources one can find a wide range of examples to a number of different problems existing even today.

Besides the processing and using of the standard technical literature the personal experiences and professional practice are also important. This means both my personal experiences and practice and those of my colleagues'.

---

4 Ullman J. - Widom J.: Adatbázisrendszerek - Alapvetés. Panem Könyvkiadó, Budapest, 1998.

5 Békéssy András -- Demetrovics János: Adatbázis-szerkezetek. Akadémiai Kiadó, Budapest, 2005.

Because of this I used some information I had collected from talking, querying, interviewing my colleagues. These kinds of information sources might have been of more importance but I had to take into account that time is money and I couldn't divide up the time of my colleagues.

I planned and executed a student questionnaire on their experiences and expectations on the scholar information system they usually used. As a teacher I was able to end up more than 91% proportion in getting valuable answers.

Besides the literature of the discipline in question I needed to apply the results of other fields of science, so my work has an interdisciplinary approach which is close to my original degree (Eötvös Loránd University, Department of General Technics) and my character.

Using and applying the results of both the general theory and the practice of systems was indispensable in my research.

I needed to search for and find the laws in force in some fields to review them at least roughly which means a little excursion onto the fields of another science. Because of the attributes of our age this is not only useful but sometimes it is indispensable as well.

Last but not least I would like to mention the fields of planning and executing measurements and experiments as an old friend. I made my dissertation in this field in 1990 at Eötvös Loránd University. I had to take into account this field because of my train of thought stands on this ground.

4. Results of the research, Summary and propositions

My first and most important result is that tthe three level data modelling has elementary importance, because the quality of an information system mainly depends on the quality of the data model. It is necessary but not enough to have a good quality end product. Because of this it is necessary to define exactly the main ideas and categories because the usage and interpretation of them is not unambiguous even today. Using accurately defined ideas is very important not only because it is a general scientific expectation but because of being the first and most important step in data modelling as well.

My second result is that although the development methods, standards and protocols are important and necessary tools in information system development but they are not enough alone to achieve a good quality. We can have the best, detailed development method which takes into account even the need of correct communication between the developer, the manager and the end user, which takes into account the importance of making proper documentations if we haven't got a good quality data model the quality of the end product will be or is being poor.

The third result is that the rational and efficient solution is to combine the three level data modelling of good quality with one of the development methods. Even the SSADM is good for that in spite of the circumstance that it seems to be underplayed. Each development method can be, or better to say, must be amended with the missing steps of the indispensable three level data modelling.

The fourth result is that I planned and executed a measurement to prove the possibility of the above mentioned solution. The result of this measurement clearly shows the differences in efficiency and loadability caused by the differences in the data models.

5. Utilization of the results

The method I showed in a specific example has a general scope, can and should be applied to any database planning task.

Each development method, standard and protocol ought to take into account the three level data modelling of good quality, the logical and physical planning based on the conceptual data model.

Based on the experiences it is necessary to work up the theory and the practice of data modelling in its full depth according to the basic work of dr. Halassy.

It is possible and necessary to combine the results into the training of database management and planning in higher education. This needs to adapt the existing objects, object programmes and requirements on one hand and to develop or to adapt the corresponding workbooks on the other.

A CASE-tool ought to be developed which would support the real three level data modelling applying all the results of the theory. It would also be needed to have a database management system which would fit for all the expectations of the relational theory of database management.

To utilize the data riches would result in a number of benefits, but I do not know any utilization of this kind, unfortunately. The detailed discussion of these problems, however, are out of the bounds of this study.

6. Publications

1989 D. Major Klára (szerk.) - prof. dr. Szücs Ervin (a Technika szerkesztőbizottság elnöke): Képes Diáklexikon - Technika. Közgazdasági és Jogi Könyviadó, Budapest, 1990. Egyes szócikkek.

1991 Gép testben gép lélek. Számítástechnika tanterv és jegyzet a Batthyány Kázmér Gimnázium és Közgazdasági Szakközépiskola, Szigetszentmiklós számára

1999 Kadocsa – Krén – Keszthelyi - Michelberger: A vállalati információs rendszerek fejlesztése a BDMF és a Ganz Holding együttműködésében. In: Gép – A Gépipari Tudományos Egyesület műszaki, vállalkozási, befektetési, értékesítési, kutatás-fejlesztési, piaci információs folyóirata, 1999. 7. szám, pp. 41-45.

2002 Keszthelyi-Michelberger: Integrált vállalatirányítási információs rendszerek gyakorlati ismertetése a felsőoktatásban. In: Arató Péter - Herdon Miklós (szerk.): Informatika a felsőoktatásban, Debrecen, 2002., p. 76. (+CD)

2002 Észrevételek az Informatika - Részletes érettségi vizsgakövetelmény és vizsgaleírás c. munkaanyaghoz (BMF főiskolai felkérés)

2002 Lektorálás: Michelberger Pál: Információs rendszerek II. jegyzetéről (A Budapesti Műszaki Főiskola felkérésére)

2002 Lektorálás: B. Virághalmy Lea: Vállalatirányítási információs rendszerek II. jegyzetéről (A Budapesti Műszaki Főiskola felkérésére)

2003 Mentette már az ehetit? In: Linuxvilág, 2003. 8. sz., pp. 60-61.

2004 Cost Effective Management with Network Resources. In: 2nd International Conference on Management, Enterprise and Benchmarking, BMF, Budapest, 2004., pp. 151-158.

2004 Korlátozható-e az internethasználat könyvtárakban és iskolákban? In: Tudományos és Műszaki Tájékoztatás, 2004. 9. sz., pp. 379-384.

2005 Information Management in the Higher Education -- The Role and Importance of the Different Technologies . In: 3rd International Conference on Management, Enterprise and Benchmarking, Budapest Tech, Budapest, 2005., pp. 123-128.

2005 Szükséges-e, szabad-e, lehet-e korlátozni az internethasználatot a könyvtárban? Előadás a „Virtuális tájékozódás, virtuális tájékoztatás” konferencián, Győr, Kisfaludy Károly Megyei Könyvtár, 2005. október 4.

2005 Számítógépterem hatékony karbantartása ssh-val. In: Iskolakultúra, 2005. 12. sz., pp. 106-109.

2006 Globalization and Cyber Community. In: Business Sciences – Symposium for Young Researchers, Budapest Tech, Budapest, 2006., pp. 77-90.

2007 A Cost Effective Solution in Intercomputer Communication. In: 5th International Conference on Management, Enterprise and Benchmarking, Budapesti Műszaki Főiskola, Budapest, 2007., pp. 223-231.

2007 Some Special Fields of Data Security. In: Business Sciences – Symposium for Young Researchers, BMF, Budapest, 2007., pp. 91-98. (A szekció legjobb előadásáért járó FIKUSZ-díj)

2008 Theory and Practice in the Education of Informatics. In: Karlovitz János Tibor (szerk.): First International Conference for Theory and Practice in Education – Current Issues in Education, Fürstenfeld, 2008. május 23. p. 34.

2008 Michelberger - Keszthelyi: Egy ERP rendszer a BMF műszaki menedzser képzésén. In: Pethő Attila - Herdon Miklós (szerk.): Informatika a Felsőoktatásban, Debreceni Egyetem, Debrecen, 2008. pp. 1-6.

2009 Possibilities of Database Based Optimization in the Higher Education. In: Karlovitz János Tibor (szerk.): Tanulás, tanítás, munkaerőpiac. Neveléstudományi Egyesület, Budapest, 2009., p. 48.

2009 The Role of Data Modelling in Information System Efficiency. In: Karlovitz János Tibor (szerk.): 2nd International Conference for Theory and Practice in Education – Teaching and Learning. Association of Educational Sciences, Budapest, 2009. p. 26.

2009 How to Measure an Information System's Efficiency? In: 7th Conference on Management, Enterprise and Benchmarking, Budapesti Műszaki Főiskola, Budapest, 2009., pp. 213-219.

2009 Lektorálás:. Karlovitz János Tibor (szerk.): Speciális kérdések és nézőpontok a felsőoktatásban. Neveléstudományi Egyesület, Budapest, 2009.

2009 Price, Value and Security. How to Manage a Database on sy's Own? In: Kóczy Á. László (szerk.): Proceedings of FIKUSZ'09 – Symposium for Young Researchers, Budapest Tech – Keleti Károly Faculty of Economics, Budapest, 2009., pp. 109-119. (A szekció legjobb előadásáért járó Fikusz-díj)

In press

2009 Számítógépterem hatékony üzemeltetése. Multimédia az oktatásban konferencia, Debrecen, június 24-25. A konferenciakiadvány megjelentetése késik.

2010 Database Aided Data Collection for Scientifical Purposes (Practice and Theory in Systems of Education)

2010 Remarks on the Efficiency of Information Systems (Acta Polytechnica Hungarica)

2010 Adatbázisok. Egyetemi jegyzet

Other activities

2005-től Member of the Association of Educational Sciences

2006-tól Member of the editorial board of the Practice and Theory in Systems of Education (P.T.S.E.)

2008 Member of the organizing committee of the conference „First International Conference for Theory and Practice in Education” (Fürstenfeld, Austria, 23/05/2008)

2009 Member of the organizing committee of the conference „Tanulás, tanítás, munkaerıpiac” (Békéscsaba, 22-24/01/2009)

2009 Member of the organizing committee of the conference „2nd International Conference for Theory and Practice in Education - Teaching and Learning” (Budapest, 29/05/2009)

2010 Member of the organizing committee of the conference „Oktatás, nevelés, élethossziglani tanulás” (Székesfehérvár, 15/01/2010)