Benefits of Using OLAP versus RDBMS for Data Analyses in Health Care Information Systems

Srebrenko Pešić, Tatjana Stanković, and Dragan Janković

Abstract—**In modern times, High qualities Information Systems are unconditional need in health care structure of developed countries. In our country, initiation process of such IS is at the very beginning. Despite that, we have tried to construct case "what-if" and give some answers which could be useful in further implementation of most advanced techniques in statistical calculations, analysis and decision-making processes in health care. OLAP systems have capabilities for fast and readable insight, making strong basis for the top-class business decisions. This paper describes advantages of OLAP over classic queries in relational databases which are in use in health care. As an example, OLAP system was created based on a database of Clinic of Neurology in Nis, as well as statistical data from Yearbook of Administration for Economics, Sustainable Development and Environment Protection.**

Index Terms—**OLAP in medicine, Health Care Information System.**

I. Introduction

INFORMATION systems are realized to help efficiency and consistency in business processes. After a while of IS existence a large amount of valid data is usually collected. In large companies or government institutes there is often a need for making some serious decisions, important for further business strategies, and based on large-scale data analysis. It is necessary to provide tools for efficient analysis on one side, and on the other, simplicity of those analyses, so people who are not IT experts could use them, like company management for example. One special segment of data processing is processing with hypothesis establishment cause, in another word – for studying. The area of medicine and medical information systems is one such area. Medical databases are very important part of every informatics society because they are directly related to the country health status, and so they affect all other society segments, so they need to be treated with special carefulness. Students, researchers, professionals and other people use medical databases to gain some data important for their activities. Those databases are further used for medicine improvement attendance, as given services quality marks, or like confirmation of some hypotheses (about certain trends and modern way of life negative aspects).

Statistic techniques and machine learning techniques are usually applied over medical data. Complexity of those techniques fluctuates from those extremely simple like histograms, to the most complex like prediction systems are. Statistical tests have wide appliance in medical researches, because they give them a simplicity, flexibility, and reliability. Basically, most experiments are performed to discover some important medical facts, which are confirmed through statistical calculations. Statistical tests are based on the hypothesis on the statistical characteristics of the analyzed medical data. The end cause is to prove correctness of the hypothesis with big trustiness. This problem becomes more complex when user has to analyze more than few data subgroups, with different combinations of risk attributes.

In the early eighties, some new methodologies for existing databases exploring have been developed. One of them is OLAP (online Analytical Processing [1]). Another definition that gives better description of the approach is Fast Analysis of Shared Multidimensional Information (FASMI) [2].

In OLAP system, users try to gain interesting but unexpected results by analyzing data subsets aggregated on different levels. OLAP techniques can have qualitative appliance in medical area, because they are intuitive, reasonable, and efficient, and on the other side again, they do not require advanced informatics knowledge from end-user. In OLAP systems, the biggest part of calculations is based on simple aggregations and counting, which are bases of the statistical tests.

Statistical methods have curtain benefits. They have simple assumptions about probability of distribution among datasets. There are no problems in such assumption when those methods are used in research with parameters which can easily be compared by querying over RDBMS. This is applicable until the moment of requirements for manipulation over data matrix. Also, statistical calculations give good results over small datasets just like over large ones. On the other side, statistical calculations have their imperfections. Basically they require great number of attempts and miss shots before they lead to some valid result. Every new attempt requires new choosing of parameters to accomplish under datasets partition.

S. Pešić is with Health Care Center, Niš, Serbia.

T.Stanković and D. Janković is with the Faculty of Electronic Engineering, University of Niš, Niš, Serbia.

One of the key benefits that OLAP system has comparing to statistical calculations is fast interactive querying through multidimensional and hierarchically organized data. Also, OLAP can be used for efficient reporting, quality control of given services, and for data integrity checking. The only deficiency of all is consumption of time needed for data warehousing.

This paper's goal is to present OLAP capabilities in Health Care systems. The contribution that OLAP systems can give to health services is not limited to one area - it refers to simplified decision-making (for management) or better tracking of medical parameters, such is frequency of some diagnosis referred to patient's age, gender, territorial partitioning, etc. Data structures and methods used in OLAP systems will be explained in next chapters. The results of applying OLAP over real medical data, and some prognosis related to larger Health Care systems will be presented after. At the end we will present summarized results and some directives for such system further developing.

II. Data and Methodology

Online Analytical Processing (OLAP) systems are very efficient tool used in complex Management Information Systems (MIS). These systems resolve next problems:

- Data analyzing according to parameters that user evaluates to be important.
- Reporting that requires exceptions or aggregations related to key indicators, trends, comparing by periods or territories, and other similar analyses.
- Business reports that require summing, exceptions and trends over different subjects.

The most important characteristic of OLAP system is multidimensional data which facilitates moving through data over "dimensions" and "measures". OLAP translates existing data from relational schemas by assigning key indicators (measures) to adequate contest (dimensions). When data is placed in multidimensional database (cube), all measures are easily and quickly available. The relation between dimensions and measures can be presented by star schema. The simplest schema presents tables with dimensions surrounding the table with the main data that comprises measures, called fact table. This is presented in Fig. 1. Table in which measures are comprised, comprise relations to dimensions also (foreign keys to outside tables).

Additional important characteristics of OLAP system are embedded and programmable analytic possibilities, and different options for data presenting and reporting. OLAP algorithms run over large datasets, and their result which overcame by using simple grouping and aggregating functions is unknown in advance.

To use OLAP system, previously we need to develop adequate multidimensional database. That process consists of standard steps [3] from which the most important and most demanded are data filtration and data importing to curtain dedicated OLAP tools. Data filtration (PREPARATION) implies fault elimination (irregular inserts, duplicates, data inconsistency, and violation of referential integrity). This step can significantly slow down OLAP system developing. Data importing problem has technical nature, because multidimensional databases can very often overcome system hardware limits. The largeness of cube can be decreased by data aggregation before transporting to OLAP system. Besides, problem with available space can be determined by dividing cube to two or more overlapping multidimensional hyper cubes (analyses are divided to several sub-analysis).

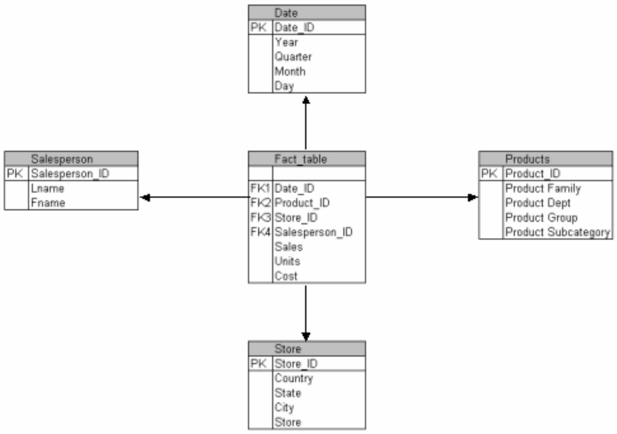

Fig. 1. Star schema example.

Health Care has experienced great improvement lately by bringing computers in many clinical and administrative processes. Yet, there is no full potentiality of using medical data as management and diagnostic decision-making source. This paper describes OLAP capacity developing process from on-line transaction processing (OLTP) system (electronic health record [4]).

The Electronic Health Record (EHR) or the Electronic Patient Record (EPR) is on-line transaction processing (OLTP) system that enables on-line inserting and updating of given health care services and documentation, medical results tracking, and real-time deciding support. Because it contains information details about daily activities, such OLTP system has great OLAP capabilities in medical, financial, and administrative area. Health Care employers have understood lately the benefits of those systems, and have been beginning to show curtain interest in data analysis which would have helped them to easily achieve answers to number of every day situation questions. Unfortunately, the most part of classical OLTP EHR systems has not suitable support for OLAP systems.

Very often there is a situation that data in HER systems are not well structured. Also is not rare that there are well structured medical information systems, but analyses that would bring to useful conclusions are not system-integrated, or cannot be performed in a real-time.

According to all above, the challenging is to:

- Choose the right data that can be important.

- Choose the data that are relevant to analyzing in a contest of easiness for later using.
- Make the connections between seemingly separated data placed in different subgroups of the system.
- Analyze great deal of data through more different parameters.
- Come to conclusions in the analysis of data, which indicate the dependence of consequential and causal connection.

OLTP systems based on a traditional RDBMS without OLAP support are not convenient for performing those tasks because much more hardware resources are needed to accomplish same results, while response time for every day usage is too slow. Therefore, to gain multidimensional system suitable for easy manipulation above datasets, we need to pass curtain phases [5]. Those phases are:

- Multidimensional model creating (determining measures, dimensions and schemas),
- Extracting, transformation, and storing data to created schemas,
- Creating and manipulating with reporting by using relational or multidimensional sources, and
- Generating information from system by using created reports (algorithm).

In our paper we used star schema for OLAP multidimensional cube. Basic (fact) table is surrounded with dimension-tables, as is presented in Fig. 2. We have created OLAP system with several measures and more then several dimensions, based on the available database of Clinic for Neurology of Clinical Center Nis. Database contains patients records collected from the beginning of year 1996. until the end of year 2008. Database migration from MS Access 2000 to MS SQL Server 2005 platform was performed. After successfully finished first step (data transformation) we have got star-schema that was suitable for developing OLAP (Fig. 2).

In an effort to establish analytics related to the possibility, need, benefits of using OLAP in public health, as well as the existence of reasons for the necessity of OLAP in a close future, we selected all possible parameters that were able to represent the measures, and for the dimensions we chose different types of patient population: gender, occupation, age, marital status, as well as doctors and diagnoses as shown in Fig. 3.

III. Study and Actual Results

OLAP is built over Clinic of Neurology Nis database, in Mucrosoft Business Intelligent Developement Studio 2005. Analytics has been done in this software package, and in ProClarity Desktop Proffessionall 6.2 during september-november 2008 period of time. Star-shema fact table has been reduced to 27000 records after significant data transformation. Those record contains informations about patients health examinations, their hospitalisations, dehospitalisations, deaths, etc. The basic idea was to come to the conclusions related to the use of OLAP in public health in general, and over the data from all Residential Health Clinic in the region. There is a plan for using such system in Health Center Nis, after the introduction of information systems and data collection period of at least a year.

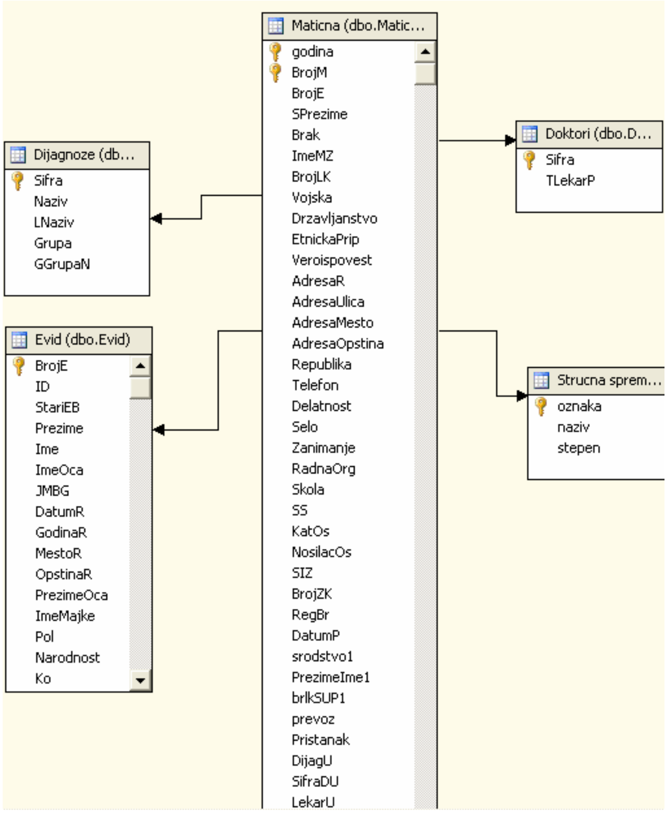

Fig. 2. Clinic of Neurology Nis star-schema.

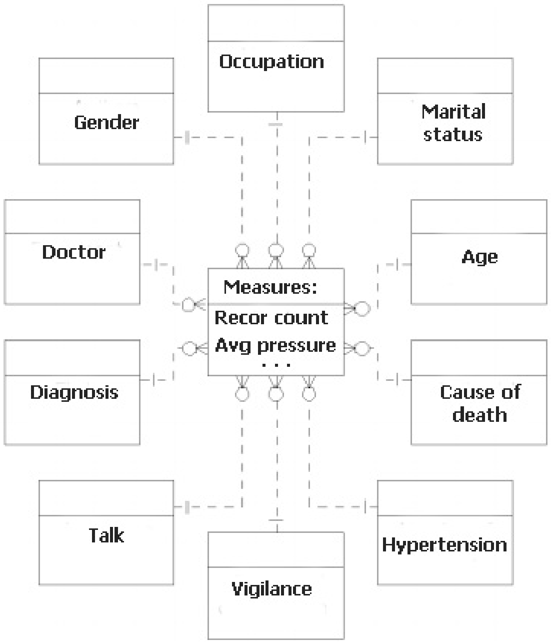

Fig. 3. Clinic of Neurology Nis hypercube model, developed for analysis on the necessity of establishing OLAP in public health.

The results obtained in the built-in OLAP showed the following:

- Cube processing time is not of importance. Cube processing on database on server cofiguration (Intel Pentium Dual CPU E2160 1.80GHz, 3.00GB RAM) lasted from 10 to 15 seconds, depending on the number of dymensions included to cube;
- Data analytics related to diagnoses, such as for example, most diagnoses that emerged in more than 400 patient examinations in the period of 10 years, rare diagnoses and doctors that establish them, diagnosis in relation to age, degree, sex and marital status of patients and similar, proved to be very simple for the end users. With the help of tools that were used for analysis (Microsoft ProClarity Desktop Profesionall), and system users without training to work with computers could very easily acquire the analysis of OLAP, only if filed names (dymensions and measures) were concise and understandable for the end users. According to that, more in database design should take into account the naming of objects and attributes.
- During the report analyses, we came very quickly to expected, but what was even more important by us, to completely UNEXPECTED results. Example: analytics of the number of patient treathments by gender, marital status and diagnosis, unexpectedly showed that there were significantly more treatments of men who are married, but of all other population – as is shown in Fig. 4. Good material for the neurologist to do research on the topic: whether married men are the most endangered population in terms of neurological?
- The time needed for OLAP quering is significantly less then the time needed for quering relational database to get the same results. For executing query that gives results (15395 records) about number of patient threatments by gender, marital status and diagnosys for frequent diagnoses on relational database, server needed ~ 7 seconds. At the same server, the time required to obtain the same results on developed cube was ~ 0.2 seconds.

Based on statistical data [4], we may be able to make the assessment for this system implementation to quantitatively greater volume of data. To get started, as the test center will be taken Health Center Nis as one of the largest institutions of its kind in the Balkans. So it will be interesting to compare the time of execution of queries – comparing the results from The Clinic for Neurology and the valuation for Health Center Nis or all health centers in the country. For this, it is necessary first to ensure the introduction of information systems in public health and their use in a given period in order to collect relevant data for the full research. However, lets look some of the statistical data that our public health has collected for years, even wothout information system. These data are presented in the Statistical Yearbook for the city of Nis for the year 2007 [6], and they are related only to the General Medicine Service (Table I).

In this spreadsheet can be seen that the number of visits to ambulance (only for the service of general medicine) a year is between 600 000 and 800 000. Observed for all primary health

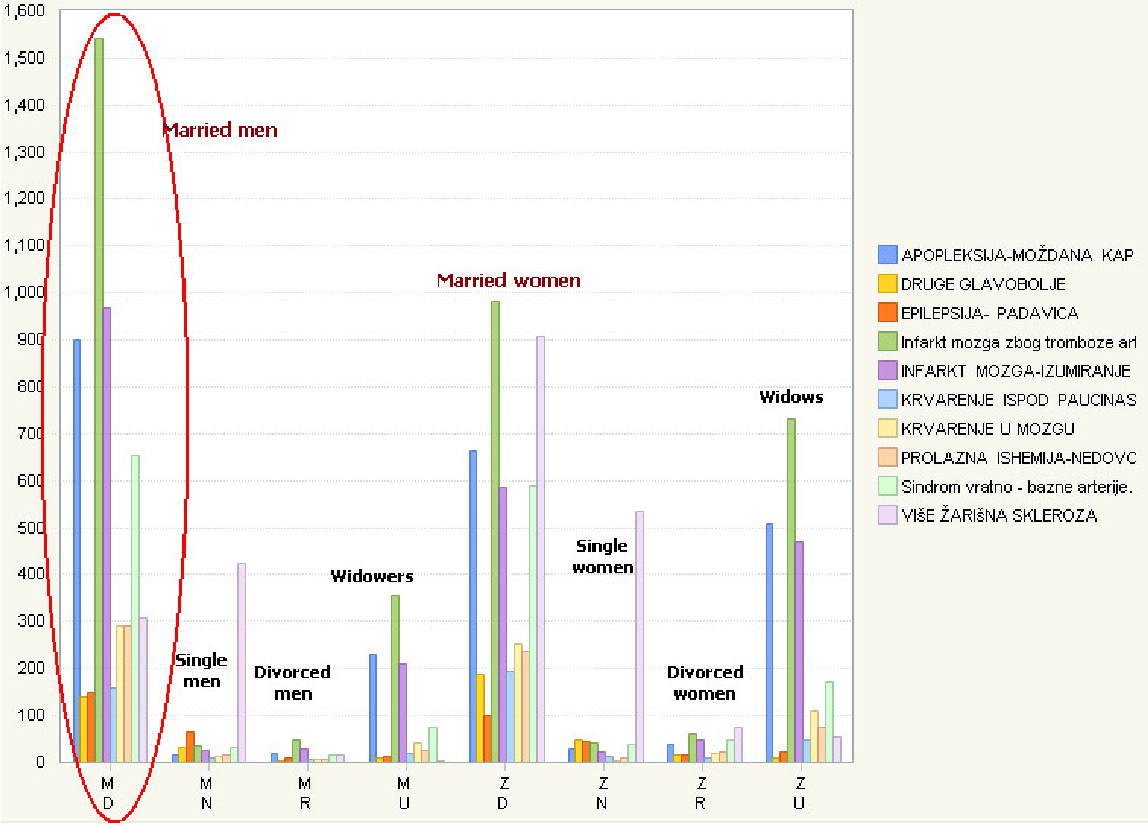

Fig. 4. Surprise factor - get unexpected results by analyzing OLAP for common diagnoses.

TABLE I
Clipping from the Table 19.7. General Medicine Service – SGN2007

Year	Threatments		**Total threatments**	Threat. per doctor	Home threatments
	First thr.	Repeated threatment			
1998	220 551	385 475	**606 026**	7 390	17 715
1999	214 549	392 159	**606 708**	7 399	17 987
2000	261 378	465 199	**726 577**	8 146	18 429
2001	278 694	507 511	**786 205**	8 276	19 613
2002	288 092	454 697	**742 789**	7 902	19 811
2003	262 603	513 943	**776 546**	8 261	20 268
2004	287 352	486 403	**773 755**	7 661	12 138
2005	275 923	532 314	**808 237**	8 164	5 069
2006	268 735	536 795	**805 530**	7 897	7 662
2007	227 938	515 049	**742 987**	6 694	17 915

care in this area this number may be up to 7-8 million per year. As for the base at which we have built OLAP system for analysis, the number of visits included is not greater then 30000. Even in this case, we have received a significant difference by comparing the time needed for quering common relational database, and OLAP cube. The time required to execute the same queries over OLAP database is 35 to 100 times less then over RDBMS. Let us mention that this database still has a bunch of textual key columns. Considering all these facts, there is a logical thought: to put our public health in a situation to apply the modern pro-European health structures, OLAP technology for analysis and business decision-making will not only have advantages over the traditional report, but will in a way be necessary.

IV. Conclusion

A conclusion section is not required. Although a conclusion may review the main points of the paper, do not replicate the abstract as the conclusion. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

The paper discussed the possibility of OLAP use in the analysis of medical data. OLAP system is implemented at the Clinic of Neurology in Nis, which has for 10 years had a total of 30,000 patient treatments. This system has enabled a quick overview of cumulative data and fast execution of complex queries. Else, such queries would not be possible over the classical relational base, or would be far slower over it. OLAP so now offers a new view of the data that have been collected for longer period of time. According to the data from the Statistical Yearbook for city of Nis, from which we have presented only few in this paper for the illustration, you may find the cost-effectiveness of this approach when applied to larger systems.

Cost-effectiveness of OLAP-for small and medium databases is questionable and must be considered for each case separately. The main question that is raised is: "Is the required time for the creation of OLAP systems worth the potential gains?“.

Today almost every RDBMS and statistical software packages include OLAP support (SQL Server, Oracle...), which is a sign that further development will go in the direction of more massive application of OLAP and appropriate techniques? In this way, access to data and the analysis of the data is provided to the experts of all different profiles that are not IT professionals. In order to use these techniques, it is necessary to pay attention to the way how databases are designed. Some aspects of this problem are presented in the paper.

References

[1] E. F. Codd, S. B. Codd, C. T. Salley, Beyond decision support, Computerworld, 27, pp. 87-90, 1993.

[2] N. Pendse, What is OLAP, http://www.olapreport.com/ fasmi.htm.

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smith, Advances in Knowledge Discovery and Data mining, MIT Press, pp. 1-34, Cambridge, 1996.

[4] Rajković, P., Janković, D., "Electronic Patient Record as a Basis of Medical Information System", XXXIX International Scientific Conference on Information, Communication and Energy Systems and Technologies ICEST 2004, Bitola, Macedonia, June 2004.

[5] G. T. Monaco, An Introduction to OLAP in SQL Server 2005, http://www.devx.com/dbzone/Article/21410/.

[6] Uprava za privredu, održivi razvoj i zaštitu životne sredine, Statistički godišnjak grada Niša 2007., pp. 199-217, Niš, novembar 2008.