PROCESSING OF LARGE DATA SETS: EVOLUTION, OPPORTUNITIES AND CHALLENGES*

Ivanka Valova, ICSR, Bulgarian Academy of Sciences, Sofia, Bulgaria
Monique Noirhomme- Fraiture, Institut d'Informatique, FUNDP, Namur, Belgium

Abstract

In the paper are analyzed the applied and theoretical results achieved, as well as some existing drawbacks in technologies for processing of large data sets-OLAP (On-line Analytic Processing), DM (Data Mining) and SDA (Symbolic Data Analysis). A comparative analysis is proposed on different types of data processing and are highlighted the pros and cons of each one of them. Here are discussed benefits and drawbacks at using of data aggregates and visualization of large data set. Some topics of interest are shown for the purposes of additional scientific study, being specifically oriented to software applications.

INTRODUCTION

Increased computer power combined with the need to analyze huge data sets, created conditions for development of new techniques and technologies. These include development of new algorithms and new approaches, e.g. use of Intuitionistic Fuzzy Logic (IFL) as well as development of new methodologies (SDA) or applying of new approaches to existing algorithms. Possession of large data base by any company is insignificant, if end users may not easily synthesize necessary information. Frequently, data has valuable additional hidden potential. This is completely new information that is displayed in the form of meaningful interrelationships between the data that are either too well hidden or too complex to be discovered just by looking at the data. The data is frequently originated from different sources. Data is retrieved, consolidated, managed and prepared for analysis. A need is identified for reliable tools for analysis of company data, which to complement the existing data management systems and to be robust enough to predict and facilitate complex analysis of business data. The goal to extract new knowledge from huge data sets has urged the need to extend standard data analysis methods (exploratory, clustering, factorial analysis, discrimination,..). A task for the research community will be also to investigate how the achievements in a specific field may be applied in new data processing technologies. It is necessary existing concepts to be clarified, the knowledge transfer to be performed in order to have a clear understanding on future research and needs to be met by developments to come. This paper is structured as follows: First are highlighted key benefits and drawbacks of different types of OLAP architectures. Second, the attention is focused on DM and different data processing methods. Then, are described different visualization methods of data sets in aggregated form. Here are discussed benefits and drawbacks at using of data aggregates. At the end application of new methodologies for analytical data processing is discussed.

*Work supported by Local Organizing Committee of PCaPAC

OLAP

The term OLAP (On-Line Analytical Processing) was initially introduced by renowned and respected database researcher E. F. Codd [1]. In 1995 he added six new rules to the original 12 for OLAP systems and restructured it into four groups, calling them "features" – Table 1. The most of OLAP software products meet the Codd's requirements for OLAP compatibility, provided that we make distinction between the rules having relation with the research approaches and application technologies. We think that some of Codd's rules may be used for improvement of the existing software technologies, and the remaining part thereof should be further developed and improved by the research community before to be proposed for practical realization. Our brief evaluation of some of these rules is given below: ***F1-***All modern software products comply with this indication. In the field of research a special attention should be drawn on the used terminology. ***F2***, ***F3, F4, F6, F7, F8-*** The experts in software technologies should exert more efforts for achievement of these requirements for on-line analytical data processing. ***F5-***The scientific research should be focused on issues for clarification of different types of models, which may be introduced. The use of mathematical calculations and definitions should be more understandable and well presented. For us the issues on terminology are remaining as subject to further discussion. For example: Multidimensional analysis or Analysis of multidimensional models? Multidimensional modeling or Modeling of multidimensional models? Dimensions or D-structures? **F13-F15** – There exist good achievements in the field of commercial products, such as Panorama technology, used in Microsoft Analysis Services. The product SAP BW meets these rules in sufficient extent. **F16-F18 –** good rules, which will be analyzed and evaluated in other paper. The basic core in the systems for analytical data processing is the creation of multidimensional models. The multidimensionality is the main requirement of Dr. Codd in the formulation of the OLAP term and in determination which software products are compatible with OLAP. The concept of a dimensional data model that could be represented in a relational database is described by Ralph Kimball [2]. This concept gained popularity and soon at the market appeared software products with relational OLAP (ROLAP) architecture. To combine the benefits of both

technologies there are increasing efforts to integrate them in new software architecture, so called hybrid OLAP (HOLAP) systems. Hybrid OLAP architecture may be defined as a system which supports and integrates multidimensional and relational storage for data in an equivalent manner in order to benefit from advantages of both technologies - MOLAP and ROLAP.

Table 1: The Codd's rules for OLAP

B Basic Features	F1-Multidimensional Conceptual View	F2-Intuitive Data Manipulation
	F3- Accessibility: OLAP as a Mediator	F4-Batch Extraction vs Interpretive
	F5-OLAP Analysis Models	F6-Client Server Architecture
	F7-Transparency	F8- Multi-User Support
S Special Features	F9-Treatment of Non-Normalized Data	F10-Storing OLAP Results: Keeping Them Separate from Source Data
	F11- Extraction of Missing Values	F12- Treatment of Missing Values
R Reporting Features	F13-Flexible Reporting	F14-Uniform Reporting Performance
	F15-Automatic Adjustment of Physical Level	
D Dimension	F16- Generic Dimensionality	F17-Unlimited Dimensions & Aggregation Levels
	F18-Unrestricted Cross-Dimensional Operations.	

Existing problem issues related with OLAP technologies may be classified as follows: There exist no generally accepted technology and terminology on issues related with OLAP systems; High prices of software products and requirements for powerful hardware; HOLAP – architectural reality or marketing hype? All vendors and developers of OLAP products are working to make their products marketable as "hybrid". It is critically important to closely examine the architectures of these new software applications, as their "HOLAP" claims may be more marketing hype than architectural reality.

OLAP is a part of scope of tools, supporting decision making process. Traditional tools for queries and reports describe what is contained in a database. Software product OLAP answers the question why certain things are true. User creates a hypothesis of specific relationship and checks its adequacy through series of queries to available data. The whole set of functions, necessary for retrieval, processing and formatting of such data is provided by OLAP processor. Efficiency of resultant query is a dominant factor for general assessment of the system. For the purposes of further scientific research of great importance are issues related with finding of most effective way of presentation of queries and balancing of dynamic load between database where the data is stored and OLAP engine. OLAP-analysis in its nature is a deductive process. Or, as Codd says, it is a process where the user "peels back one or more layers of the onion through subsequent simple queries". Processes related with DM methods differ from those in OLAP systems, as instead to verify validity of hypothetic models, they use available data to create new models. Fig. 1 shows relation between business need in certain organization and using of OLAP&DM&SDA&IFL.

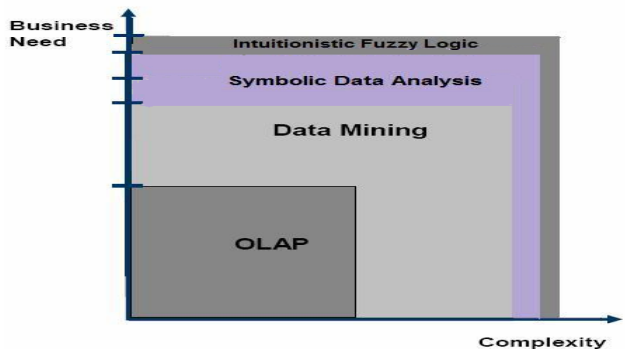

Figure 1: OLAP&DM

DATA MINING (DM)

DM – deriving of valid, previously unknown information from large databases and using it at taking of critical business decisions. Data analysis is performed aiming to discover hidden or not presumed earlier connections and patterns in analyzed data. DM uses methods from different fields, such as statistics, neural networks, machine learning, etc [2]. Most frequently used DM methods, being realized in modern software products are as follows: *Decision tree* - here data is displayed divided into categories. The resulting model is presented in the form of a tree structure; *Clustering* - divides data up into homogeneous groups; *ABC Analysis (Pareto analysis)*- The entire production is divided into three groups of products – A, B and C. Group A products contribute to highest annual earnings, and these of Group C – to the lowest earnings. Pareto analysis is a predecessor of ABC classification for determination of value of inventory stock for each group of products. Pareto (1848-1923), discovered the law on nonproportional causes, where are. 80% of consequences are result of 20% of their causes. Pareto analysis helps to highlight most important issues (these ones, called by Juran "the vital few") and all efforts to be focused on its solving. For finding of solution of a complex issue and removal of causes of the same, as well as for search of appropriate measures different methods may be used. An appropriate method for categorization of already identified causes is the "Ishikava diagram", named after Japanese professor Kaoru Ishikava. Such diagram has a complex branched structure and is called also fishbone diagram; *Association analysis (affinity analysis or Market Basket Analysis (MBA))* is designed to determine associations between different events. MBA is an algorithm that examines a long list of transactions in order to determine which items are most frequently purchased together. The main problems in using of algorithms of MBA occur in analysis of product, which is sold only 1 or 2 times as reflected in analyzed data file. Products, which do not attract commercial interest, should

not be included in the file, as the algorithm will determine rules including products, which are not statistically significant. It is preferable all products to be equally represented. Unlike the decision tree classification, in clustering and association analysis, the models are determined on the basis of the data itself. To obtain maximum effect, users must use such methods that are most suitable for a certain organization.

VISUALIZATION OF DATA

Tools for graphic presentation and visualization are important help engines for data preparation and their importance in terms of data analysis is not to be underestimated. Visual analysis allows the discovery of overall trends but also smaller hidden patterns. Models, links and missing values are frequently perceived easier, when displayed graphically, than if presented as list of figures or text. Several taxonomies and surveys are available for data visualization. Use of pie chart on Fig. 2

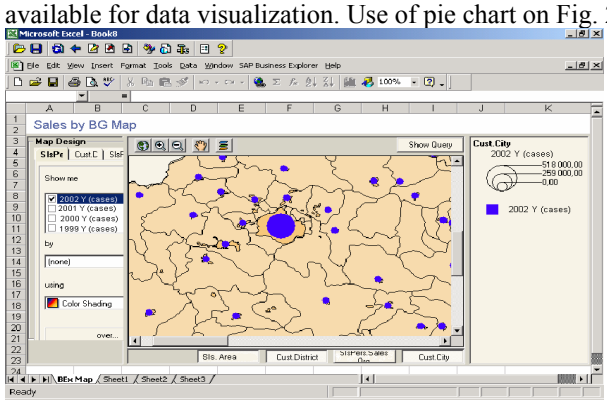

Figure 2: Map presentation using pie chart -The size of circles in individual regions shows different volume of sales of certain goods.

allows us to obtain a full picture on sales within the territory of given country. In case we are interested in sales dynamics of several stocks within certain timeframe, Fig. 3 is our ideal case. This representation has been used to visualize specific symbolic objects varying with time and is a development by BE scientists. The user is provided with several means of interactivity. They concern standard visualization features such as zoom, rotation of figures and setting of colour and fonts [3]. This representation has been used to visualize a symbolic object varying with time. What these both methods of

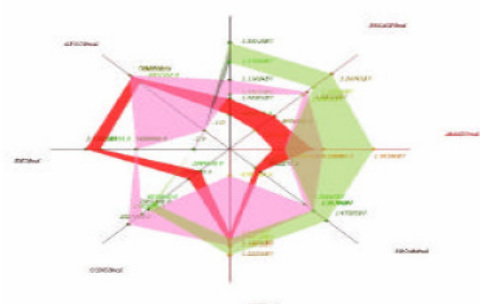

Figure 3: Example of superposition of stars. 8 stocks value for three different weeks.

representation have in common is the visualization of data sets on aggregated form. Such sets may be represented as existing data or as results of preliminary analysis.

Pros and cons in use of aggregates: Aggregates improve performance at runtime of certain query, but increase loading time; Aggregate must be checked regularly whether additional data is missing or not. When to be compressed associated aggregates – upon entering of data or after the data was already loaded in database? Aggregates allow fast access to data in reporting mode.

APPLICATION OF NEW METHODOLOGIES

Symbolic Data Analysis (SDA)

The French scientist Edwin Diday defines "Symbolic Data Analysis" (SDA) as the extension of standard Data Analysis [4]. The data descriptions of the units are called "symbolic" when they are more complex than the standard ones due to the fact that they contain internal variation and are structured. Symbolic data happen from many sources in order to summarize huge sets of data. Symbolic data analysis has been developed to solve the problem of the analysis of data known on an aggregated form, i.e. where quantitative variables are given by intervals and where categorical variables are given by histograms. SDA in its essence was intended as methodology and in this regard the contributions are indisputable. Recent information about new SDA modules is available on www.assoproject.be/

Intuitionistic Fuzzy Logic (IFL)

Intuitionistic Fuzzy (IF) Logic can be used in evaluation of the models for large data set. IF Set is defined as follows: $A=\{\langle x, \mu_A(x), \nu_A(x)\rangle / x\in E\}$, where E is fixed set, functions $\mu_A:E\rightarrow[0,1]$ and $\nu_A:E\rightarrow[0,1]$ give degree of membership and non-membership of the element $x\in E$ to set A. Set A is subset to E and $\forall x\in E$: $0\leq\mu_A(x)+\nu_A(x)\leq 1$. Value $\pi_A(x)=1-\mu_A(x)-\nu_A(x)$ gives the degree of non-determinacy of the element x:E to the set A.

CONCLUSIONS

The quality of strategic and business decisions, being taken by using of the new technologies discussed in this paper, is significantly higher and they are much well-timed as compared with the decisions taken by using of traditional methods.

REFERENCES

[1] Codd E., S. B. Codd, C. T. Salley. Providing OLAP to user-analysts: An IT mandate. Technical report, 1993.
[2] Kimball R., Data Warehouse Tool Kits. John Wiley & Sons, Toronto, 1996.
[3] M. Noirhomme-Fraiture: Visualization of Large Data Sets: the Zoom Star Solution, Journal of Symbolic Data Analysis, vol 1, July 2002. http://www.jsda.unina2.it
[4] E. Diday, Introduction a e'Approche Symbolique en Analyse des Donnees. Premiere Jouneles Symbolique-Numerique, CEREMADE, Universite Paris - Dauphine, 1987, 21-56.