*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

***Work supported by Local Organizing Committee of PCaPAC 2008**

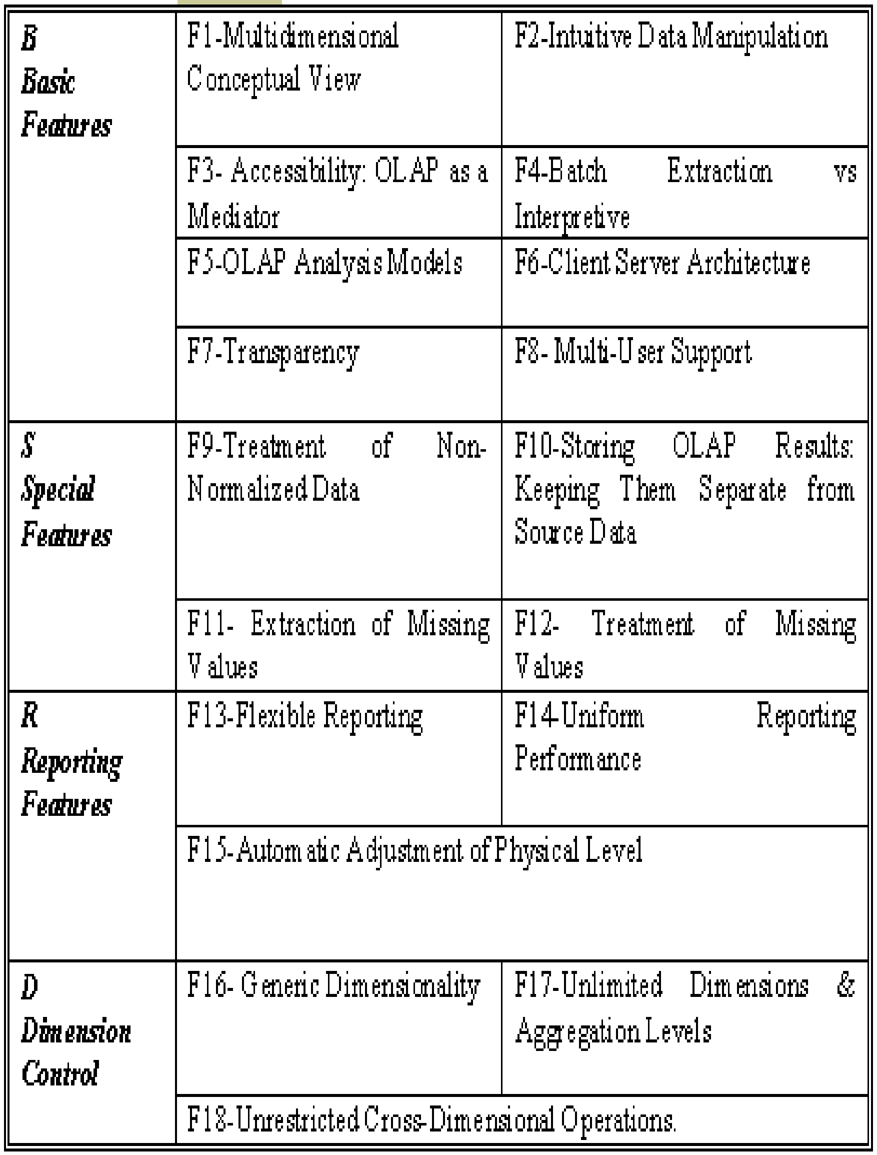
RESUME

This article discusses modern technologies and concepts for processing of large data sets.

- *OLAP (On-line Analytic Processing)*
- *DM (Data Mining)*
- *SDA (Symbolic Data Analysis)*
- *IFL (Intuitionistic Fuzzy Logic)*

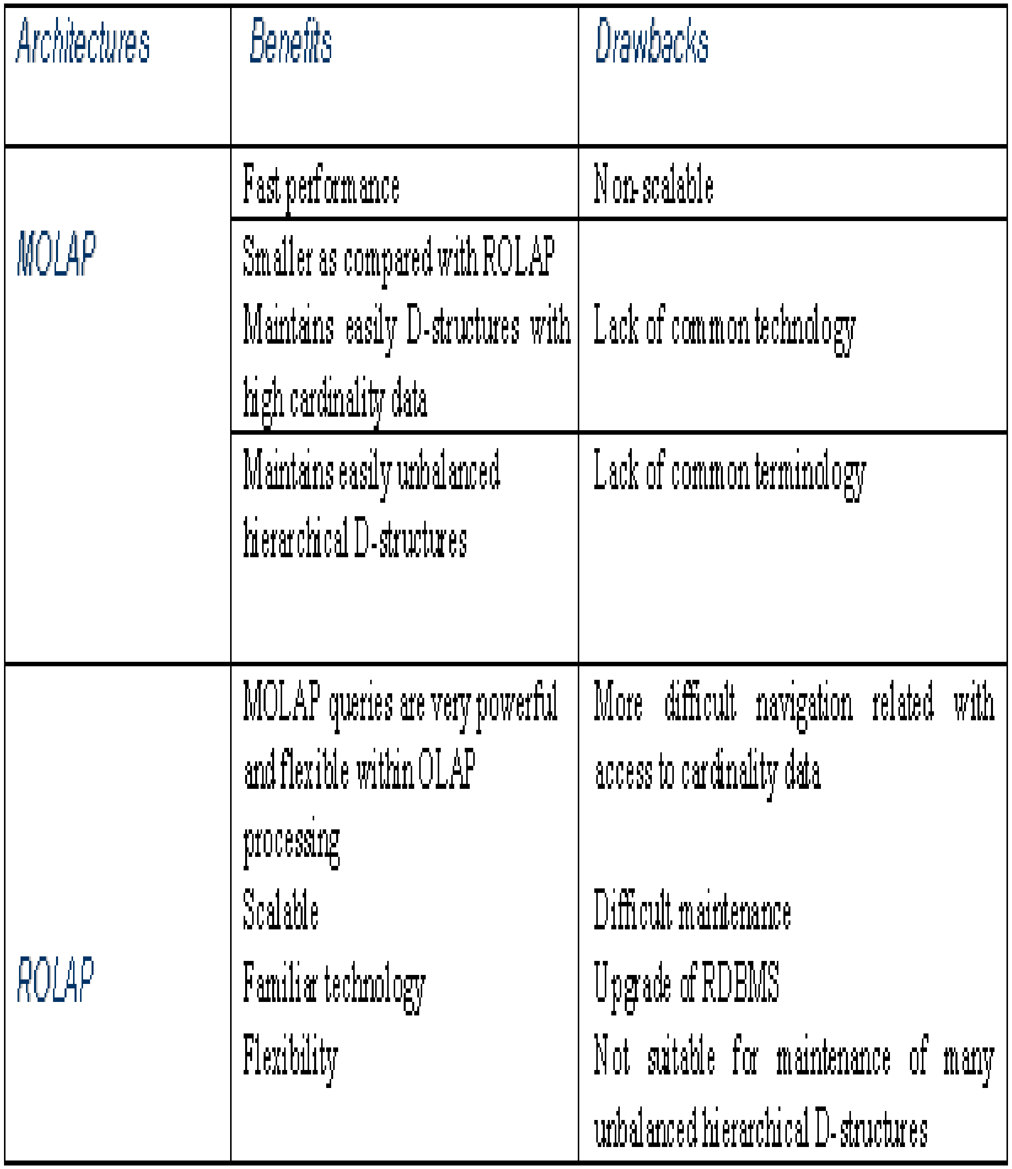
OLAP

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 Control*	F16- Generic Dimensionality	F17-Unlimited Dimensions & Aggregation Levels
	F18-Unrestricted Cross-Dimensional Operations.	

Comparative analysis – MOLAP and ROLAP



Architectures	Benefits	Drawbacks
MOLAP	Fast performance	Non-scalable
	Smaller as compared with ROLAP Maintains easily D-structures with high cardinality data	Lack of common technology
	Maintains easily unbalanced hierarchical D-structures	Lack of common terminology
ROLAP	MOLAP queries are very powerful and flexible within OLAP processing Scalable Familiar technology Flexibility	More difficult navigation related with access to cardinality data Difficult maintenance Upgrade of RDBMS Not suitable for maintenance of many unbalanced hierarchical D-structures

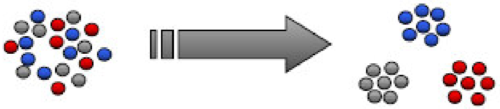
Data Mining (DM)

DM – deriving of valid, previously unknown information from large databases and using it at taking of critical business decisions.

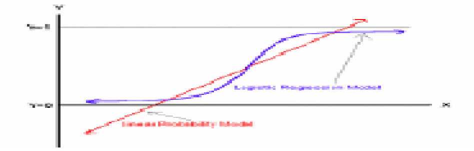
Most frequently used DM methods, being realized in modern software products are as follows:

- *3.1Decision tree*

- *3.2 Clustering*



- *3.3 Weighted Score tables and Regression (linear regression and nonlinear regression)*



- *3.4 ABC Analysis (Pareto analysis)*

- *3.5. Association analysis (affinity analysis or Market Basket Analysis (MBA))*

To obtain maximum effect, users must use such methods that are most suitable for a certain organization.

Visualization of aggregated 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.

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.

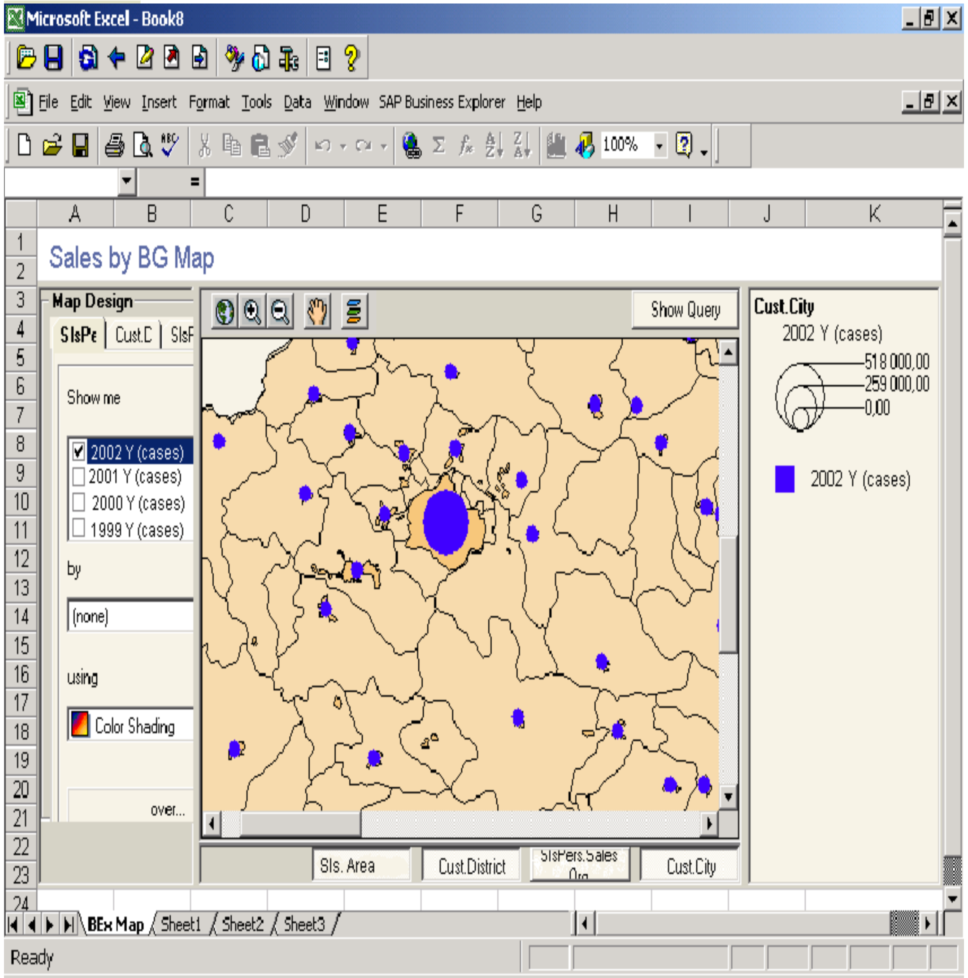
•When to be compressed associated aggregates – upon entering of data or after the data was already loaded in database?

•Aggregates allow fast access to data in reporting mode.

Visualization of aggregated data

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

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



Sales by BG Map

Show me

☑ 2002 Y (cases)
☐ 2001 Y (cases)
☐ 2000 Y (cases)
☐ 1999 Y (cases)

by

(none)

using

Color Shading

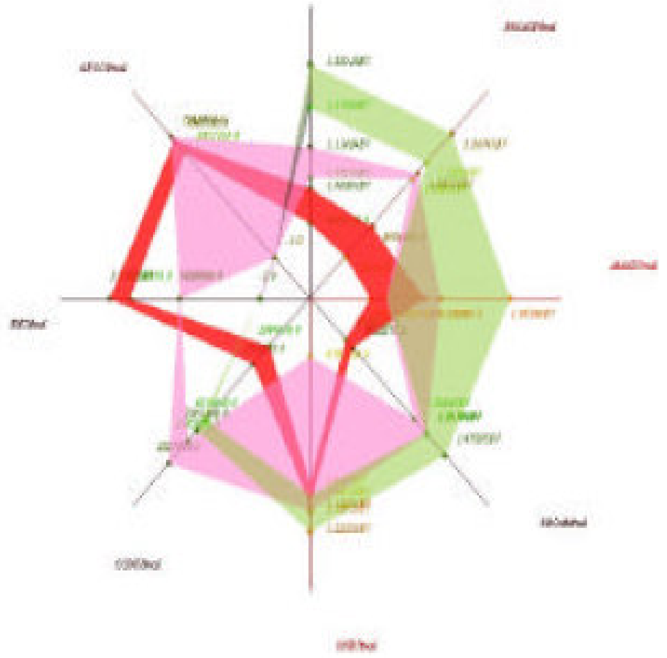
over...

Cust.City

2002 Y (cases)

518 000,00
259 000,00
0,00

🟦 2002 Y (cases)



This representation has been used to visualize a symbolic object varying with time.

Application of new methodologies- SDA and IFL

SDA

The French scientist Edwin Diday defines "Symbolic Data Analysis" (SDA) as the extension of standard Data Analysis. 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.

Intuitionistic Fuzzy Logic (IFL)

IFL can be used in evaluation of the models for large data set. IF Set is defined as follow:

$$A=\{\langle x, \mu A(x), \nu A(x)\rangle / x\in E\},$$

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.