БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ . BULGARIAN ACADEMY OF SCIENCES

КИБЕРНЕТИКА И ИНФОРМАЦИОННИ ТЕХНОЛОГИИ · Том **4**, № 1
CYBERNETICS AND INFORMATION TECHNOLOGIES · Volume **4**, No 1

София . 2004 . Sofia

Overview of Research and Software Approaches for Multidimensional Data Analysis

Ivanka Valova[1], *Bozhan Zhechev*[2], *Vladimir Valov*[2]

[1] *Institute of Control and System Research,1113 Sofia*
E-mail: vania@icsr.bas.bg
[2] *Institute of Computer and Communication Systems,1113 Sofia*
E-mail: jechev@iac.bg

Abstract: *In this paper a comparative analysis was made of the most frequently quoted research multidimensional models and approaches, specifying advantages and disadvantages of each one. The application software products are also studied. The E. Codd's rules are analyzed and it was evaluated its applicability in the available software products for analytical data processing. A special focus is made on the multidimensionality, being a key requirement in discussed models, used in OLAP systems.*

Keywords: *fact, D-structures, relational algebra, software products for on-line processing*

1. Introduction

In 1993 E. Codd formulated the requirements for on-line multidimensional analytical processing. He presented 12 rules, which are now well known (and available for download from vendors' Web sites). They were followed by another six rules in 1995. The rules for online processing are: *F1*–Multidimensional Conceptual View, *F2* – Intuitive Data Manipulation, Accessibility:*F3* – OLAP as a Mediator, *F4*-Batch Extraction vs Interpretive, *F5* – OLAP Analysis Models, *F6* – Client Server Architecture, *F7*– Transparency, *F8* – Multi-User Support, *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, *F13* – Flexible Reporting, *F14* – Uniform Reporting Performance, *F15* – Automatic Adjustment of

Physical Level, *F16* – Generic Dimensionality, *F17* –Unlimited Dimensions & Aggregation Levels, *F18* –Unrestricted Cross-dimensional Operations.

In this type of modeling the information is divided into *facts* and *dimensions (D-structures)*.

Fact (F) – it represents the data, subject to analysis. The facts contain numerical attributes. The fact in a multidimensional scheme is the object, which contains *measures.* (Measures evaluate attributes of fact).

D-structure *(Dimension*) – different initial viewpoints at data selection, which will be used during fact analyzing. *D-structures* contain mainly description attributes.

During the process of preparation of the Bulgarian version of this paper we have discussed thoroughly the translation of the term Dimension. In our working variants we accepted the terms "*Dimension* and *Dimensia-Izmerenie"*, but its usage in numerous cases created conflicts with accepted terminology in field of mathematics. Therefore, we accepted the notion *D-structure* as a translation of the English term *Dimension,* and at the same time we do not underestimate the fact that the term *Dimension* is widespread and used by the computer experts and exists in the model developed by us according to the requirements of the used software product.

D-structure *represents a connected directed graph, provided that each peak of the graph corresponds to a given aggregation level, and its arcs reflect "part-whole" relations between the objects within the aggregation levels. The above definition and OLAP terminology in general will be a subject of our future paper.*

Analysis of multidimensional models – the possibility data (*F*) to be united, displayed and analyzed according to multiple *D-structures* (*D*), in ways, which are meaningful for one or more specific corporate and scientific analysts in any moment of time.

Multidimensional model – Presentation of OLAP data in the form of a cube (referred also as *infocube* or *hypercube*) with data or in the form of *a "star"* type scheme (referred as multidimensional scheme), by the use of facts and a set of *D-structures*, based on the notion of hierarchy of *D-structures*.

The Infocube (abbr. the cube) may be presented as a limited space of Cartesian *n*-functionally dependent levels of accumulation, to a set of cells in class (C_c).

Very often, in analytical data processing, the metaphor *cube of data* is used. *Such notion was accepted, nevertheless that from a viewpoint of mathematics the derived figure is not always a cube.* The notions *infocube* or *hypercube* are also used. Each cell in this cube represents an intersection point of different types of *D-structures* – D_i at exact defined level and participates in the determination of quantitative indicators of the fact.

In the following sections, some research and market works in the field of online multidimensional processing and data warehousing systems are summarized. Section 2 briefly presents the software products for on-line multidimensional processing. In section 3 we present the comparative analysis of research multidimensional models. Our conclusion is given in Section 4.

2.Software products for multidimensional processing

The software database vendors, which embraced multidimensional analysis, are Oracle, IBM, Microsoft, Arbor Software Corporation, MicroStrategy, Lotus, CA, Accure Software, Sybase, etc. All developing or marketing products in this area. Fig. 1 presents some important events in this direction.

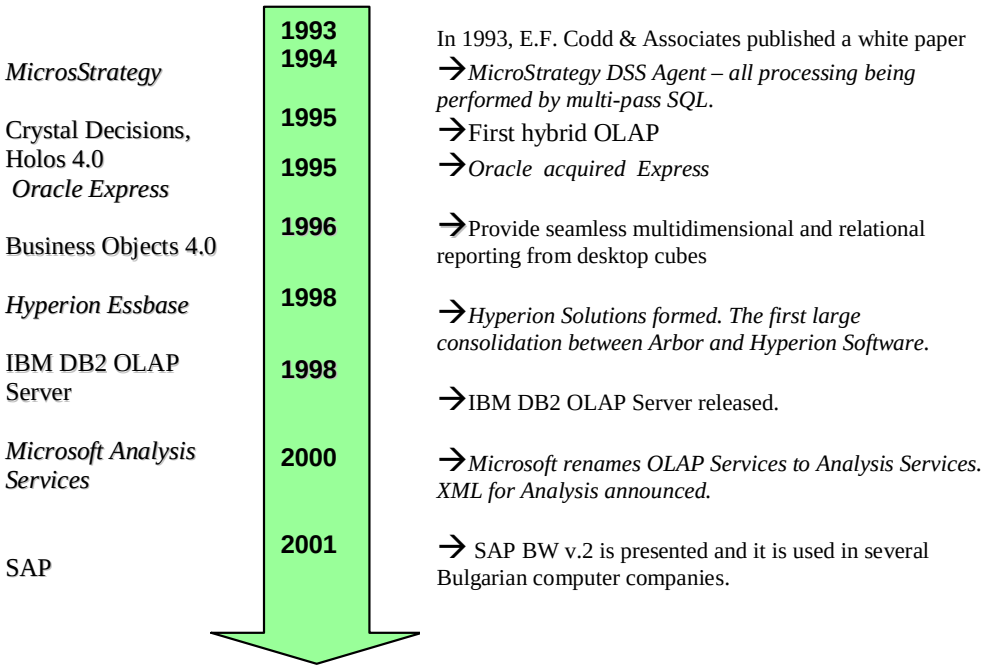

Fig. 1. Software products for multidimensional processing

Most of OLAP software products presented in Fig.1 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 and it will be a subject of a more detailed overview in another paper.

F1 – all software products presented in Fig.1 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 (it is valid also for authors of this paper). 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 another paper.

The software products for real time data processing can be classified as follows:

Hypercube products – This approach is used by *Essbase* (Arbor Software), *Hyperian Enterprise, CLIME, Comshare FDC*. This is a single-cube logical structure. Data is entered for every combination of dimension members and all parts of the data space have identical dimensionality. Vendors of these products emphasize their greater simplicity to the end-user [2].

Multi-cube products – in this approach developers segment the database into a set of D- structures each of which is composed by a subset of the overall number of dimensions. An example of the multi-cube approach is SAP BW (Business Information Warehouse). Fig. 2 shows the architecture of SAP Business Information Warehouse (BW). BW uses a multi-level architecture to provide the maximum degree of flexibility. BW can extract and use data provided by a variety of sources [7]. These include R/3 and R/2 systems, non-SAP systems, flat files, commercial data providers and even other BW systems.

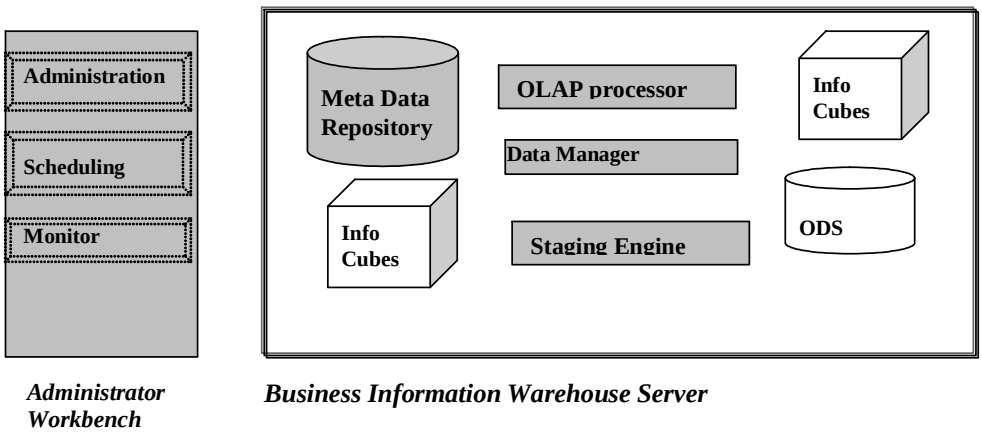

Fig. 2. Architecture of SAP BW

The BW server provides all the necessary tools to modeling, extracting, storing and accessing the data. Since the description of the data, regardless of source, is contained in a common meta data repository, data from a variety of sources can be combined for enhanced analysis possibilities [5].

The core of BW 2.0 data warehouse construction and administration is the Administrator Workbench. This is a software tool for constructing multidimensional InfoCubes to support multidimensional analysis. BW is the hub for other SAP components such as R/3, Customer Relationship Management (CRM), and Strategic Enterprise Manager (SEM) [7]. The Business Information Warehouse operational data store (ODS) creates two types of database tables for each ODS object: for active use and as a working copy to prepare new data loads. Once the new data has been verified, the changes are activated and everyone sees a consistent data view. The term objects here are not limited to a physical database table, but includes all methods, transformations, or intra-object communication rules and workflow that make up an entity to support business activities [5].

Research MDM Terminology	BW Terminology
Fact	Key Figure
Dimension Attributes	Characteristic / Navigational Attribute / Reporting Attribute / (external) Hierarchy
Dimensions (D-structures)	Dimension Data / Master Data / Text Data / Hierarchy Data /(SID Data)

Fig. 3. Terminology

Fig. 3 presents terminology used by research community and software developers of SAP product – Business Information Warehouse. The market survey has proven the need for further research activities in this area.

3. Research approaches

3.1. Multidimensional model based on relation algebra

One of the first multidimensional data model and one of the most referenced is the model of Agrawal discussed in [8]. The center of their approach is relational algebra and operations which can be translated to SQL. The main features of this model are the following:

– The data is organized in a multidimensional cube. All cell values can either be an n-tuple or from the set {0, 1}. A cell containing "1" means that this combination of dimension values exists. An n-tuple represents the existence of a record with n measures and a "0" marks cells with no contents. The dimensions have no structure or order and the elements are addressed by their name.

– Symmetric treatment to not only all dimensions but also to measures. Support for multiple hierarchies along each dimension and support for adhoc aggregates.

A multidimensional cube *C* is formally defined as

(*D*, *E*(*C*), *N*), where:

D is a set of *k* dimension names. Each dimension has a domain dom_i.

E(*C*) is a function mapping $dom_1 x \ldots x\, dom_k$ to an *n*-tuple (the cell values of the cube *C*) or to {0, 1}.

N is an *n*-tuple containing the names of the members of the *n*-tuples contained in the cube.

Example1. For instance, if it is necessary to analyze company sales, we could do it attending to three dimensions, i.e Date (when something was sold), Store (where it was sold), Product (what was sold):

C= (*D, E*(*C*), *N*),

D = {*store, day, product* },

E(*C*) contains the mapping of coordinates to 4-tuples or "0".

In the model of Agrawal, no distinction is made between measures and dimensions. In Example 1, *sales* is just another dimension. Thus, a cube in this model may have more logical dimensions than the number of dimensions used to physically store the cube in a multidimensional storage system.

The proposed operators are minimal. Each one of them is defined on the cube and produces as output a new cube. For example the push operation converts dimensions into elements and can be expressed as follow:

Input: ***C***, ***Di***.

Output: ***Ca***

push (*C, D i)= Ca.*

E(Ca)(d1, …., dk) = ***g*** $\oplus$ ***di***

where ***g*** = ***E(Ca)(d1, …, dk)***.

$\oplus \rightarrow$ **0**, if ***g*** **= 0,**

$\oplus \rightarrow$ < ***di*** > *if* ***g*** **= 1**

In other cases $\oplus$ connected **g** and < ***di*** >.

The proposed operators are: *Pull, Destroy dimension, Restriction, Join, Merge.*

The model does not present static construct dimension levels. All of the structural and functional information has to be included in the query.

Structured measures can be expressed easily by n-tuples, which are cube elements. Derived measures can be expressed by using a self-join operation on the cube. In this case the definition of the calculation has to be given in the query.

Example 2.

destroy*_dim(*
merge*(*
restric*t(*
restric*t(C,* $f_{r\text{-}up}$*(day→year)(day) =2003), region(store)=GD),*
{[store, $f_{r\text{-}up}$ *(store →store_type)],*
[day, $f_{r\text{-}up}$ *(day →month)]],*
[product, $f_{r\text{-}up}$ *product →all)]}, favg),product)))*

The expressive power of the model is powerful as the relational algebra as the relational operators union, intersect and difference can be expressed using the basic operator set.

3.2. Star schema approach

In books [2] and [3,4] some multidimensional design patterns are presented. The models of Kimball are composed by a central fact table and a set of smaller dimension ones surrounding it – a star schema. The fact table contains measures (numerical values) and dimensional tables contain attributes (textual characteristics).

Some authors argue that it is also important to normalize schemas (also known as "snowflaking" – Fig. 4.). As a side effect, it shows aggregation hierarchies in the dimensions. However, the saved space is irrelevant while query performance is worsened.

The problems with changes in the data along time in practical multidimensional model are discussed as "Slowly Changing Dimensions". The old values must be kept, because the facts previous to the change are still related to them, while new ones will be referred by the fact occurring from now on [4].

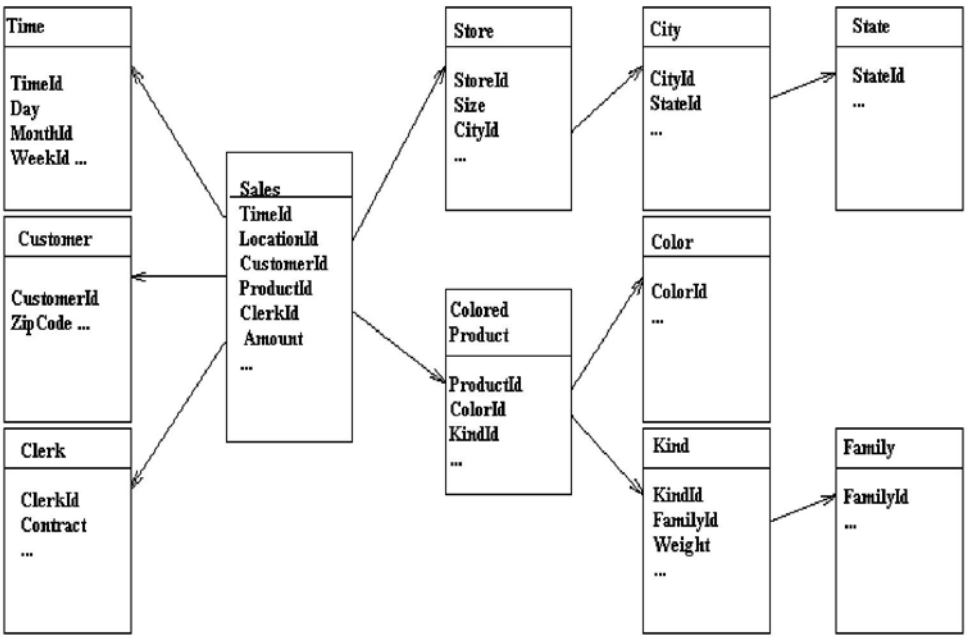

Fig. 4. An example of a snowflake schema [2]

The book of K i m b a l l [4] is for designers, managers, and owners of data warehouse and for research workers. Working models of all the databases described in the book [3] are included in CD-ROM.

3.3. An approach based on grouping algebra

The grouping algebra provides a declarative approach to multidimensional analysis. Basic concept is a multidimensional cube consisting of a number of relations, dimensions, and for each combination of dimension tuples, an associated (scalar) data value representing a single fact attribute.

A multidimensional cube is a set of dimension relations ri and a mapping from a n-dimensional tuple (coordinate) to a scalar value.

pair (F, m) where $F=\{(D_1,r_1), ..., (D_n,r_n)\}$

D_i, $i = 1,..., n$ is the dimension names, R_i are sets of attribute names.

r_i, $i = 1,..., n$ – a relation on R_i for each i and m is a mapping from $\{\{(D_1, t_1), ... ,(D_n,t_n)\}| \forall\ 1 \le i \le\ n{:}t_i \in r_i\}$ to V (a set of scalar values).

Cubes in the same multidimensional database share dimansion relations. This means that, if two cubes have the same dimension name, they are using the same dimension relation.

Approaches	Presentation of multidimensional model	Advantages	Disadvantages
Agrawal, Gupta, and Sarawagi [1]	(*D*, *E*(*C*), *N*)	As the *n*-tuples are allowed in the capacity of cube's elements, the recorded structural measures may be expressed easier	The model does not contain information on *D-structures (Dimensions)*. There is no static structure representing levels *L*. The complete structural and functional information should be included in the query
Li, Wang [5]	(*F*, *?*), *F={(D1,r1), …., (Dn, rn)},* $? \rightarrow \{(D_i,t_i) / 1 \leq t \leq n \ \& \ t_i \in r_i\}$ *to V*	The grouping algebra may be considered as an extension of the relational algebra The proposed algebra provides an independent, declarative approach of performance according to the requirements of the analysis of multidimensional models	It is not possible to visualize complex measures. A possible solution is to construct a separate cube for each attribute of measure. The derived measures should be calculated separately

4. Conclusion

As the comparison shows each of research and software models has its specific advantage but none of them is good enough. This makes a combination of the approaches desirable in future. Each approach presents its own view of multidimensional analysis requirements, terminology and formalism. Consequently, there is no commonly accepted formal multidimensional data model established. Such a model is necessary as a basis for an accepted standardized logical data model. This would allow practitioners and researchers to specify their multidimensional data models in a unified way.

References

1. Agrawal, R., A. Gupta, S. Sarawagi. Modeling Multidimensional Databases. – In: Proc. of the 13th ICDE, Birmingham, U.K., April 1997, 232-243.
2. Inmon, W. H. Building the Data Warehouse. Second edition, John Wiley & Sons, 1996.
3. Kimball, R. The Data Warehouse Tool Kits. John Wiley & Sons, 1996.
4. Kimball, R. The Data Warehouse Lifecycle Toolkit. John Wiley & Sons, 1998.
5. Li, C., X. Wang. A data model for supporting on-line analytical processing. – CIKM, 1996, Hong Kong, 53-58.
6. Oracle Corporation. Frequently Asked Questions about Oracle Express Objects and Oracle Express Analyzer.

http://otn.oracle.com/products/express/
7. SAP Business Information Warehouse.
http://www.sap-ag.de/solutions/industry
8. V a l o v a, I., B. Z h e c h e v. Data warehouse models for real time processing. – Automatics and Informatics'03, Bulgaria, Sofia, 6-8 October, Vol.1, 2003, 37-40

Преглед на изследователските и софтуерните подходи за многомерен анализ на данни

Иванка Валова[1], *Божан Жечев*[2], *Владимир Валов*[2]

[1] *Институт по управление и системен анализ, 1113 София*
E-mail: vania@icsr.bas.bg
[2] *Институт по компютърни и комуникационни системи, 1113 София*
E-email: jechev@iac.bg

(Р е з ю м е)

В тази статия е направен сравнителен анализ на най-често цитираните изследователски многомерни модели и подходи чрез посочване на предимствата и недостатъците на всеки от тях. Изследвани са и приложните софтуерни разработки. Анализирани са правилата на Е. Код и е оценено тяхната приложимост в съвременните софтуерни продукти за аналитична обработка на даннни. Специално внимание е обърнато на многомерността, като ключово изискване в разглежданите модели, предназначени за OLAP системите.