A Survey on Logical Models for OLAP Databases

Panos Vassiliadis, Timos Sellis
National Technical University of Athens
Department of Electrical and Computer Engineering
Computer Science Division
Knowledge and Database Systems Laboratory
Zografou 15773, Athens, Greece
{pvassil,timos}@dbnet.ece.ntua.gr
Tel: +301-772-1602, Fax: +301-772-1442

1 *Introduction*

The database world has always divided the modeling tasks based on three different perspectives: the *conceptual* one, dealing with the high level representation of the world, the *physical* one, dealing with the details of the representation of the information in the hardware, and the *logical* one, which acts as an intermediate between the two aforementioned extremes, trying to balance a storage-independent paradigm and a natural representation of the information in terms of computer-oriented concepts.

OLAP databases could not escape from this rule; several conceptual (e.g. [BaSa98],[Kimb96]) and physical (e.g. [Sara97]) models exist; yet, in the sequel we will focus on the presentation of different proposals for *multidimensional data cubes*, which are the basic *logical* model for OLAP applications.

It has been argued that traditional relational data models are *in principle* not powerful enough for data warehouse applications, and that data cubes provide the functionality needed for summarizing, viewing, and consolidating the information available in data warehouses. Despite this consensus on the central role of multidimensional data cubes, and the variety of the proposals made by researchers, there is little agreement on finding a common terminology and semantic foundations for a data model.

We have proceeded in the following categorization of the work in the field: on the one hand there are the commercial tools -which actually initiated the work on the field; we present them first, along with terminology and standards, in Section 2. On the other hand there are the academic efforts, which are mainly divided in two classes: the relational model extensions and the cube-oriented approaches. We present the former in Section 3 and the latter in Section 4. In Section 5, we attempt to a comparative analysis of the various efforts and finally, in Section 6 we present concluding remarks and pending research issues.

2 *Terminology, Products and Standards*

2.1 Terminology

A good definition of the term OLAP is found in [OLAP97a]: "…On-Line Analytical Processing (OLAP) is a category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that has been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user. OLAP functionality is characterized by dynamic multidimensional analysis of consolidated enterprise data supporting end user analytical and navigational activities including calculations and modeling applied across dimensions, through hierarchies and/or across members, trend analysis over sequential time periods, slicing subsets for on-screen viewing, drill-down to deeper levels of consolidation, rotation to new dimensional comparisons in the viewing area etc. …". A standard terminology for OLAP is provided by the OLAP Council [OLAP97a].

The focus of OLAP tools is to provide multidimensional analysis to the underlying information. To achieve this goal, these tools employ multidimensional models for the storage and presentation of data. Data are organized in *cubes* (or *hypercubes*), which are defined over a multidimensional space, consisting of several dimensions. Each dimension comprises of a set of aggregation levels. Typical OLAP operations include the aggregation or de-aggregation of information (*roll-up* and *drill-down*) along a dimension, the *selection* of specific parts of a cube and the re-orientation of the multidimensional view of the data on the screen (*pivoting).*

2.2 Products and Technologies

The debate on the underlying physical model, supporting OLAP, is centered around two major views. Whereas some vendors, especially vendors of traditional relational database systems (RDBMS), propose the *ROLAP architecture* (Relational On-Line Analytical Processing) [MStr95, MStr97, Info97, RedB97], others support the *MOLAP architecture* (Multidimensional On-Line Analytical Processing) [Arbo96]. The advantage of the MOLAP architecture is, that it provides a direct multidimensional view of the data whereas the ROLAP architecture is just a multidimensional interface to relational data. On the other hand, the ROLAP architecture has two advantages: (a) it can be easily integrated into other existing relational information systems, and (b) relational data can be stored more efficiently than multidimensional data.

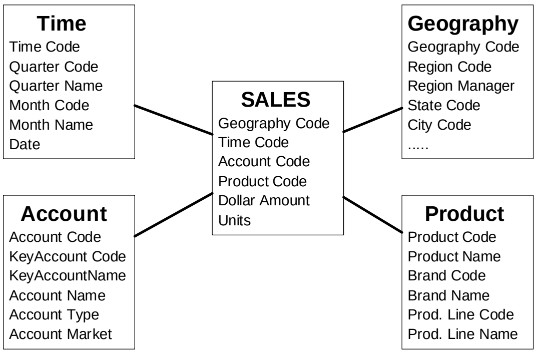

Figure 1. Star schema [Stan96]

In a ROLAP architecture, data are organized in a *star* (Figure 1) or *snowflake* schema. A star schema consists of one central *fact* table and several denormalized *dimension* tables. The *measures* of interest for OLAP are stored in the fact table (e.g. Dollar Amount, Units in the table SALES). For each dimension of the multidimensional model there exists a dimension table (e.g. Geography, Product, Time, Account) with all the levels of aggregation and the extra properties of these levels. The normalized version of a star schema is a snowflake schema, where each level of aggregation has its own dimension table.

Multidimensional database systems (MDBMS) store data in n-dimensional arrays. Each dimension of the array represents the respective dimension of the cube. The contents of the array are the measure(s) of the cube. MDBMS require the precomputation of all possible aggregations: thus they are often more performant than traditional RDBMS [Coll96], but more difficult to update and administer.

2.3 Benchmarks and Standards

The OLAP Council has come up with the *APB-1 benchmark* [OLAP97b] for OLAP databases. The APB-1 benchmark simulates a realistic OLAP business situation that exercises server-based software. The standard defines a set of dimensions with respect to their logical perspective. The logical database structure is made up of six dimensions: *time*, *scenario*, *measure*, *product*, *customer*, and *channel*. The benchmark does not assume a specific underlying physical model: the input data are provided in the form of ASCII files. The operations nicely simulate the standard OLAP operations and include bulk and incremental loading of data from internal or external data sources, aggregation or drill-down of data along hierarchies, calculation of new data based on business models, etc.

The *TPC-D benchmark* [TPC98] models a decision support environment in which complex ad hoc business-oriented queries are submitted against a large database. TPC-D comprises of a hybrid star and snowflake schema, involving several dimension and fact tables. The benchmark is definitely relational-oriented: there is no explicit treatment of cubes and dimension hierarchies. Of course, one can always deduce them implicitly from the underlying schema; nevertheless, the dimensions seem too simple in their structure and depth. The benchmark is accompanied by a set of queries which seem to be close to the usual queries in a DSS environment. These queries do not fit the pattern of typical OLAP operations, which are sequential and interactive in their nature. Currently, a revision of the benchmark is under preparation.

The *OLEDB for OLAP* [MS98] standard has been developed by Microsoft as a set of COM objects and interfaces, destined to provide access to multidimensional data sources through OLEDB. OLEDB for OLAP employs a model for cubes and dimensions, that supports the logical notions already explained in section 2.1. Moreover, it provides a language of *MultiDimensional eXpressions* (MDX) for the calculation and presentation of cubes. OLEDB for OLAP provides a good intuition on the entities comprising a multidimensional database; nevertheless it has several disadvantages: it lacks a solid theoretical background (e.g. there is no definition of the *schema* of a multicube) and combines *presentational* with *computational* issues. The result is a complex and, to some extent, hard to use (although powerful enough) language.

The *Metadata Interchange Specification* [Meta97] was proposed by the Metadata Coalition, an open group of companies such as IBM, Sybase, Informix, etc. The Metadata Interchange Specification (MDIS) provides a standard access mechanism and a standard application programming interface to control and manage metadata with interchange specification-compliant tools. MDIS tries to present a metadata metamodel for a wide set of database models (relational, object-oriented, entity-relationship, etc.), with a model for multidimensional databases belonging to this set. The model proposed by MDIS supports the notion of dimension which just comprises from a set of levels. Cubes are not directly modeled in the MDIS model.

3 Relational Extensions

3.1 Models for OLAP

The *data cube* operator was introduced in [GBLP96]. The *data cube* operator expands a relational table, by computing the aggregations over all the possible subspaces created from the combinations of the attributes of such a relation. Practically, the introduced *CUBE* operator calculates all the marginal aggregations of the detailed data set. The value '*ALL*' is used for any attribute which does not participate in the aggregation, meaning that the result is expressed with respect to all the values of this attribute.

In [LW96] a multidimensional data model is introduced based on relational elements. Dimensions are modeled as *dimension relations*, practically annotating attributes with dimension names. Cubes are modeled as functions from the cartesian product of the dimensions to the measure and are mapped to *grouping relations* through an applicability definition. A grouping algebra is presented, extending existing relational operators and introducing new ones, such as ordering and grouping to prepare cubes for aggregations. Furthermore, a multidimensional algebra is presented, dealing with the construction and modification of cubes as well as with aggregations and joins. For example, the operator *roll* is almost a monotone roll-up. Finally, a relation can be grouped by intervals of values; the values of the “dimensions” are ordered and then "grouped by", using an auxiliary table.

In [BPT97] multidimensional databases are considered to be composed from sets of tables forming denormalized star schemata. Attribute hierarchies are modeled through the introduction of functional dependencies in the attributes of the dimension tables. Nevertheless, this work is focused on the data warehouse design optimization problem and not on the modeling of cubes or cube operations.

In [GL97] *n-dimensional* tables are defined and a relational mapping is provided through the notion of *completion*. An algebra (and an equivalent calculus) is defined with classical relational operators as well as restructuring, classification and summarization operators. The expressive power of the algebra is demonstrated through the modeling of the data cube and monotone roll-up operators.

In [GL98] a new extension of the relational model and a new language are proposed. The underlying model is an extension of the relational model to handle *federated names*. A complex name is a pair, comprising of a name (or concept) and a finite set of associated criteria set, relating the concept to a common, global set of criteria. An extension of SQL, *nD-SQL* is also provided, along with its mapping to an extension of the relational algebra. The applicability of the language to OLAP operations is shown through a set of examples, practically modeling the *CUBE* operator of [GBLP96]. The authors give different semantics to the *ROLLUP* and *DRILLDOWN* operators than the ones we give here. Moreover, results on the optimization of the execution of queries are also provided.

3.2 Relationship with Statistical Databases

A lot of relevant work has been done in the past in the area of *statistical databases* [Shos97]. In [Shos97] a comparison of work done in statistical and multidimensional databases is presented. The comparison is made with respect to application areas, conceptual modeling, data structure representation, operations, physical organization aspects and authorization/security issues. The basic conclusion of this comparison is that the two areas have a lot of overlap, with statistical databases emphasizing on conceptual modeling and OLAP emphasizing on physical organization and efficient access.

In [OOM85, OOM87] a data model for statistical databases is introduced. The model is based on *summary tables* and operators defined on them such as construction/destruction, concatenation/extraction, attribute splitting/merging and aggregation operators. Furthermore, physical organization and implementation issues are discussed. [OOM85] is very close to practical OLAP operations, although discussed in the context of summary tables.

In [RR91] a functional model ("Mefisto") is presented. Mefisto is based on the definition of a data structure, called "statistical entity" and on operations defined on it like summarization, classification, restriction and enlargement.

4 Cube-Oriented Models

There have also been efforts to model directly and more naturally multidimensional databases; we call these efforts *cube-oriented*. This does not mean that they are far from the relational paradigm – in fact all of them have mappings to it – but rather that their main entities are *cubes* and *dimensions*.

In [AGS95], a model for multidimensional databases is introduced. The model is characterized from its symmetric treatment of *dimensions* and *measures*. A set of minimal (but rather complicated) operators is also introduced dealing with the construction and destruction of cubes, join and restriction of cubes, and merging of cubes through direct dimensions. Furthermore, an SQL mapping is presented.

In [CT97], a multidimensional database is modeled through the notions of *dimensions* and *f-tables*. Dimensions are constructed from hierarchies of *dimension levels*, whereas f-tables are repositories for the factual data. Data are characterized from a set of *roll-up* functions, mapping the instances of a dimension level to instances of another dimension level. A query language is the focus of this work: a calculus for f-tables along with scalar and aggregate functions is presented, basically oriented to the formulation of aggregate queries. In [CT98a] the focus is on the modeling of multidimensional databases: the basic model remains practically the same, whereas ER modeling techniques are given for the conceptual modeling of the multidimensional database. A mapping to physical entities such as relations and multidimensional arrays is provided. In [CT98b] a graphical query language as well as an equivalent algebra is presented. The algebra is a small extension to the relational algebra, including a roll-up operator, yet no equivalence to the calculus is provided.

In [Vass98] dimensions and dimension hierarchies are explicitly modeled. Furthermore, an algebra representing the most common OLAP operations is provided. The model is based on the concept of the *basic cube* representing the cube with the most detailed information (i.e. the information at the lowest levels of the dimension hierarchies). All other cubes are calculated as expressions over the basic cubes. The algebra allows for the execution of sequences of operations as well as for drill-down operations. A relational mapping is also provided for the model, as well as a mapping to multidimensional arrays.

In [Lehn98] another model is presented, based on primary and secondary multidimensional objects. A *Primary Multidimensional Object* (PMO), which represents a cube, consists of : a cell identifier, a schema definition, a set of selections, an aggregation type (e.g. sum, avg, no-operator) and a result type. A *Secondary Multidimensional Object* (SMO) consists of all the dimension levels (also called “dimensional attributes”) to which one can roll-up or drill-down for a specific schema. Operations like Roll-up, Drill-down, Slice, Dice etc. are also presented; yet not all of them are defined at the instance level. In [LAW98], which is a sequel to the previous paper, two multidimensional normal forms are proposed, defining (a) modeling constraints for summary attributes and (b) constraints to model complex dimensional structures.

In [GJJ97] the CoDecide model is – informally – presented. The so-called *tape model* consists of structured hierarchies called *tapes* (corresponding to dimensions). Each tape consists of a set of hierarchically interrelated *tracks* (corresponding to *levels*). The intersection of tracks defines a *multidimensional matrix*. Operations like roll-up and drill-down are defined for the tape model. It is important to note that the tape model can combine several matrices, defined as networks of crossing tapes. Moreover, the tape model is a the lower part of a layered set of models, representing the logical perspective. On top of it, the *transformation*, *visualization* and *control* models are defined, belonging essentially to the *presentational* perspective.

5 Comparison

In the sequel, we present a comparison of the various models. The first list of requirements for logical cube models is found in [BSHD98]. In our approach we followed the discrimination between entities and operations and came up with three big categories of attributes for cube models. The first group of attributes deals with the representation of the multidimensional space: as usual, we check whether entities are modeled as cubes or tables (denoted by *C* or *T* respectively) and whether level hierarchies are modeled, or not. The second group of attributes deals with language issues: the character of the query language (procedural, declarative, visual), the direct support of sequences of operations and a subjective characterization of how naturally the classical OLAP operations are modeled. The third group is concerned with the existence of physical mappings to relations and/or multidimensional arrays.

In Table 1, 'SQL ext.' indicates extension of SQL, and N/A means that the information is not directly available in the material examined (papers).

		Multidimensional space		**Language aspects**					**Physical representation**	
		Cubes	*level hierarchies*	*Procedural QL*	*Declarative QL*	*Visual QL*	*Seq. of operations*	*natural repr.*	*relational mapping*	*m/d mapping*
Relational-Oriented	GBLP96	T			SQL ext.				✓	
	LW96	T	implicitly	algebra				✓	✓	
	GL97	T	✓	algebra	calculus				✓	
	GL98	T		✓	✓				✓	
	BPT97	T	✓						✓	
Cube-Oriented	AGS95	C		algebra					✓	✓
	CT97, 98, 98a	C	✓	algebra	calculus	✓		✓	✓	✓
	Vass98	C	✓	algebra			✓	✓	✓	✓
	Lehn98, LAW98	C	✓	algebra				✓	✓	implic.
	GJJ97	C	implicitly	N/A	N/A	✓		✓	N/A	N/A
Standards	APB-1	C	✓	natural lang.				✓		
	TPC-D	T			SQL				✓	
	OLEDB	C	✓	C++ calls	SQL-like			✓	✓	implic.
	MDIS	T	✓							
Statistical	OOM85	T	implicitly	algebra				✓	✓	
	RR91	T	implicitly	algebra			✓	✓	✓	

Table 1. Comparison of the various cube models.

6 Conclusions

In this paper we provided a categorization of the work in the area of OLAP logical models by surveying some major efforts, from commercial tools, benchmarks and standards, and academic efforts. We have also attempted a comparison of the various models along several dimensions, including representation and querying aspects.

Clearly, a lot of interesting work can be expected in the area. The issue of reaching a consensus on the modeling issues is still open, both in the logical and the conceptual perspective. Devising a common standard declarative language is also of high importance. Moreover, there is potential for useful results, in the area of logical optimization and caching rules (in order to exploit the possibility of reusing existing cubes for the computation of new ones), through the use of a generic logical multidimensional model (independently from the underlying physical model).

Acknowledgments. This work is supported by the European Commission in ESPRIT Long Term Research Project DWQ (Foundations of Data Warehouse Quality). We would also like to thank Prof. Matthias Jarke.

7 References

[AGS95] R. Agrawal, A. Gupta, S. Sarawagi. Modeling Multidimensional Databases. IBM Research Report, IBM Almaden Research Center, September 1995.

[Arbo96] Arbor Software Corporation. *Arbor Essbase*.http://www.arborsoft.com/essbase.html, 1996.

[BaSa98] F. Baader, U. Sattler. Description Logics with Concrete Domains and Aggregation. Proc. of the 13th European Conference on Artificial Intelligence (ECAI-98). 1998.

[BPT97] E. Baralis, S. Paraboschi, E. Teniente. Materialized View Selection in a Multidimensional Database. In 23rd VLDB Conference, Athens, August 1997

[BSHD98] M. Blaschka, C. Sapia, G. Höfling, B. Dinter. Finding your way through multidimensional data models. In 9th Intl. DEXA Workshop, Vienna, Austria, August 1998.

[Coll96] George Colliat. OLAP, Relational, and Multidimensional Database Systems. SIGMOD Record, Vol. 25, No. 3, September 1996.

[Coll96] G. Colliat. *OLAP, Relational, and Multidimensional Database Systems*. SIGMOD Record, Vol. 25, No.3, September 1996.

[CT97] L. Cabbibo, R. Torlone. Querying Multidimesional Databases. 6th DBPL Workshop, 1997

[CT98] L. Cabbibo, R. Torlone. A Logical Approach to Multidimensional Databases. In 6th EDBT, 1998.

[CT98a] L. Cabibbo, R. Torlone. From a Procedural to a Visual Query Language for OLAP. In 10th SSDBM Conference, Italy, July 1998.

[GBLP96] J. Gray, A. Bosworth, A. Layman, H. Pirahesh. Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tabs, and Sub-Totals. Proceedings of ICDE '96, New Orleans, February 1996.

[GJJ97] M. Gebhardt, M Jarke, S. Jacobs. A Toolkit for Negotiation Support Interfaces to Multidimensional Data. In Proc. of the 1997 ACM SIGMOD Conf., Arizona, USA, 1997

[GL97] M. Gyssens, L.V.S. Lakshmanan. A Foundation for Multi-Dimensional Databases. In 23rd VLDB Conference, Athens, August 1997.

[GL98] F. Gingras, L. Lakshmanan. nD-SQL: A Multi-dimensional Language for Interoperability and OLAP. Proceedings of the 24th VLDB Conference, N. York, August 1998.

[Info97] Informix, Inc.: *The INFORMIX-MetaCube Product Suite.* http://www.informix.com/informix/products/new_plo/metabro/metabro2.htm, 1997.

[Kimb96] R. Kimball. The Data Warehouse Toolkit: Practical techniques for building dimensional data warehouses. John Wiley. 1996

[LAW98] W. Lehner, J. Albrect, H. Wedekind. Normal Forms for Multidimensional Databases. In 10th SSDBM Conference, Italy, July 1998.

[Lehn98] W. Lehner. Modeling Large Scale OLAP Scenarios. In 6th EDBT, 1998.

[LS97] H. Lenz, A. Shoshani. Summarizability in OLAP and Statistical databases. In 9th SSDBM Conference, 1997.

[LW96] C. Li, X. Sean Wang. A Data Model for Supporting On-Line Analytical Processing. CIKM 1996.

[Meta97] Metadata Coalition: Meta Data Interchange Specification, (MDIS Version 1.1), August 1997, available at http://www.he.net/~metadata/standards/

[MS98] Microsoft Corp. OLEDB for OLAP February 1998. Available at http://www.microsoft.com/data/oledb/olap/

[MStr95] MicroStrategy, Inc. *Relational OLAP: An Enterprise-Wide Data Delivery Architecture.* White Paper, http://www.strategy.com/wp_a_i1.htm, 1995.

[MStr97] MicroStrategy, Inc. *MicroStrategy's 4.0 Product Line.* http://www.strategy.com/launch/4_0_arc1.htm, 1997.

[OLAP97] OLAP Council. OLAP AND OLAP Server Definitions. 1997 Available at http://www.olapcouncil.org/research/glossaryly.htm

[OLAP97a] OLAP Council. The APB-1 Benchmark. 1997. Available at http://www.olapcouncil.org/research/bmarkly.htm

[OOM85] G. Ozsoyoglu, M. Ozsoyoglu, F. Mata. A Language and a Physical Organization Technique for Summary Tables. In SIGMOD Conference, Austin, Texas, May 1985.

[OOM87] G. Ozsoyoglu, M. Ozsoyoglu, V. Matos. Extending Relational Algebra and Relational Calculus with Set-Valued Attributes and Aggregation Functions. ACM TODS 12(4), 1987.

[RedB97] Red Brick Systems, Inc.. *Red Brick Warehouse 5.0.* http://www.redbrick.com/rbs-g/html/whouse50.html, 1997.

[RR91] M. Rafanelli, F.L. Ricci. A functional model for macro-databases. SIGMOD Record, March 1991, 20(1).

[Sara97] Sunita Sarawagi. Indexing OLAP Data. Data Engineering Bulletin 20(1): 36-43 (1997)

[Shos97] A. Shoshani. OLAP and Statistical Databases: Similarities and Differences. Tutorials of PODS, 1997.

[Stan96] Stanford Technology Group, Inc. *Designing the Data Warehouse on Relational Databases.* http://www.informix.com/informix/corpinfo/zines/whitpprs/stg/metacube.htm, 1996.

[TPC98] TPC: TPC Benchmark D. Transcation Processing Council. February 1998.

Available at http://www.tpc.org/dspec.html

[Vass98] P. Vassiliadis. Modeling Multidimensional Databases, Cubes and Cube Operations. In 10th SSDBM Conference, Italy, July 1998.