Executive Information Systems, Inc.

Evaluating OLAP Alternatives

By

Joseph M. Firestone, Ph.D.

White Paper No. Four

March 28, 1997

OLAP Alternatives

The rush to develop data warehouses and data marts has gained considerable momentum from the presence of server-based On-line Analytical Processing (OLAP) [1] tools, including: Multidimensional server-based (MDOLAP) tools; a number of Relational OLAP (or ROLAP) alternatives; and a new tool called Sybase IQ which uses a technology we can call Vertical Technology OLAP (VTOLAP). By now the gold rush in the OLAP marketplace is in full swing, with different products striving for both marginal differentiation and substantial technical and operational advantages over competitors.

How do we choose an OLAP product for a data warehouse? Carefully, as the old saying goes, and with respect to relevant criteria. This White Paper will provide a set of criteria for product evaluation in specific project contexts. My perspective is that of an independent consultant with no product axe to grind, and no conscious a priori preference for any of the three types of OLAP, an OLAP agnostic if you will. Before getting to the criteria, however, it is useful to briefly review the three categories of OLAP and to identify some of the more prominent products in each category.

MDOLAP Alternatives

There are four leading vendors of MDOLAP servers: Arbor Software (Essbase), Kenan Technologies (Acumate Enterprise), Oracle/IRI (Express), and D & B/Pilot Software (Lightship). In addition, a server version of Cognos PowerPlay, and a new multidimensional server from the SAS Institute have recently been released.

These and other multidimensional server tools are based on the idea that a multidimensional view of business or program data is consistent with and reflects business logic and common sense, and is much more relevant to practical decisionmaking than the table-oriented view presented by standard relational and flat file data bases. A multidimensional view of data as cubes and/or hypercubes, naturally segments data into cells lying at the intersection of causally or descriptively relevant categorical dimensions.

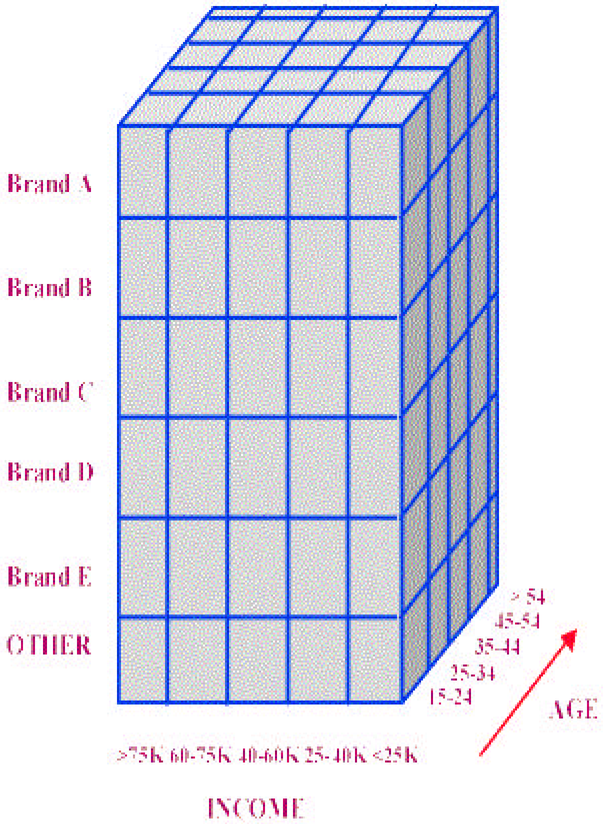
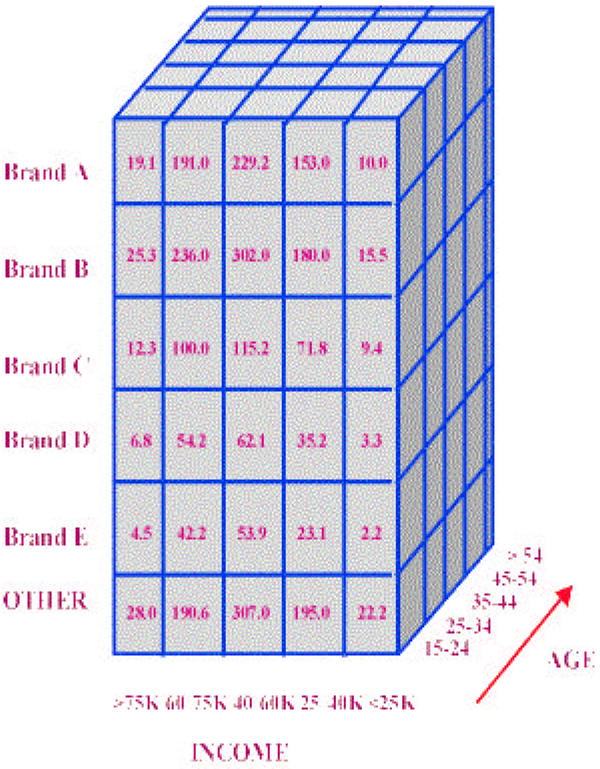


Figure One -- A Three Dimensional, Unpopulated, Array

Multidimensional database servers group field category values into dimensions, and then dimensions into multidimensional arrays. Specific field category values that may occur in data (logically possible category values) identify either the rows or columns of array dimensions, and grouped field categories identify the row or column array dimensions themselves. Figure One provides an example of a three dimensional, but unpopulated array.

Arrays are populated by placing actual field values in the cells of the array, data generally imported from flat files or relational databases. The dimensions of the array categorize the values in the cells, which, in turn, provide data variable values for the segments of the entity population defined by the categorization. The values in the cells are never values of the dimensional categories. Instead they are values of attributes that vary across the dimensions. Figure Two provides an example of a three dimensional array populated with units sold values.



	>75K	60-75K	40-60K	25-40K	<25K
Brand A	19.1	191.0	229.2	153.0	10.0
Brand B	25.3	236.0	302.0	180.0	15.5
Brand C	12.3	100.0	115.2	71.8	9.4
Brand D	6.8	54.2	62.1	35.2	3.3
Brand E	4.5	42.2	53.9	23.1	2.2
OTHER	28.0	190.6	307.0	195.0	22.2

Units Sold (in 000's) Varies Across the Three Dimensions of Age, Income, and Brand. The Numbers Shown Summarize Age Variations in Units Sold Across Brand, and Income. For Example, 19.1 Thousand Units of Brand A Were Sold to Individuals of All Ages with Income Greater than 75K.

Figure Two -- A Three Dimensional Array Populated With Units Sold Values

Attributes or variables often used to populate arrays are business measurements, such as prices, costs, sales, profits, probabilities of response to promotions, and averages of customer lifetime value. In general, values in arrays are values of continuous, or at least frequency variables, while dimensional category values are those of discrete variables, or continuous variables whose values have been categorized to accomodate the OLAP perspective.

The multidimensional array has a position in the multidimensional logical model analogous to the position of the table of field values in the relational model. Arrays provide a basic segmenting structure within which data description and analysis takes place, but to perform such analysis through query processing, it is necessary to specify higher level logical operations between components of multidimensional arrays and between multidimensional arrays themselves. Multidimensional databases can call on the full set of logical operations available to relational databases.

In addition, however, there are certain logical relations and operations that are much easier to perform in multidimensional databases because they are "hard-wired" into the design of commercial products and need not be assembled from clever and exacting manipulation of more basic logical operations. These operations include:

- defining parent-child relations between dimensions and constructing dimensional hierarchies across geography, organization, time and other important organizing concepts;
- easily performing matrix calculations that allow whole vectors or slices of

arrays to be operated on at once;
- ranging or subsetting (also known as "dicing,") multidimensional arrays to provide more focused descriptions, reports, and analyses;
- rotation (also known as "data slicing,") to examine a different view of the multidimensional array being queried without having to reassemble the array from basic data; and
- aggregating or disaggregating multidimensional arrays to diplay higher or lower levels in a dimensional hierarchy such as time period, geography, or organization (known as "rolling-up" or "drilling-down").

Multidimensional databases share with relational databases either friendly query languages or visual tools that make ad hoc querying practical for database marketing analysts, sales analysts, financial ananlysts, and other practitioners of chain querying. Multidimensional database vendors, moreover, feel that their query languages are more efficient than SQL. Many examples exist comparing SQL queries with multidimensional ones. Almost invariably multidimensional queries are a fraction of the size of SQL queries. An important factor in ease of querying is the presence of the high level operations allowing data dicing, slicing, rolling-up, drilling-down and matrix arithmetic. It is easy and highly intuitive to formulate ad hoc queries using multidimensional products, and excellent visual tools exist to aid the process of ad hoc query chaining.

In the area of query performance, multidimensional databases, unaided by indexing or special hardware, exhibit improved performance over relational databases based on E-R data models. The reason for increased performance is that multidimensional databases use the high level logical operations named earlier to either retrieve summaries or counts that immediately fulfill queries, or at a minmum, to access a place in the multidimensional database that allows query processing to proceed through scanning only a small portion of the data.

An ad hoc query like "How many blue, minivans were sold by Chevy dealers in Minnesota in 1993?" requires no scanning of a multidimensional database. The cell with the answer can be reached by drilling down to State, and dicing and slicing to the face and cell of the array which has the information in it. And if the next query is, and "how many were sold in Minneapolis/St. Paul?" the answer is even faster in coming because it only requires a "drill-down."

ROLAP Alternatives

While ROLAP vendors have entered the market only in the last few years, there are four of them that have received a good bit of attention in the trade press. Two are among the fastest growing corporations in the United States. The four companies are: Microstrategy [2], Information Advantage, Stanford Technology Group (recently acquired by Informix), and IQ Software. All offer ROLAP suites including analytical server engines, reporting and analysis tools, system design tools, and web enabling software.

ROLAP products are said to produce performance comparable to that of MDOLAP tools, but to be much more scalable to larger database sizes. ROLAP vendors adhere to the OLAP doctrine that end users should be presented with a multidimensional view of business data. But they strongly disagree that it is necessary to physically store data multidimensionally, or to use a logical model that views data as a cube or hypercube, in order to provide a multidimensional view of the data. Instead, ROLAP practitioners have developed the concepts of dimensional modeling including such data modeling concepts as the star schema, the snowflake schema, the constellation, and qualities. Figure Three illustrates a star schema with a fact table in the middle of the model and dimensional tables clustered around it.

Dimensional models clearly present multidimensional views of data. They do so by presenting a flattened view of the dimensional slices of a Decision Support system (DSS). Like a multidimensional server, ROLAP requires a separate server engine for analytical processing. But ROLAP differs from OLAP in that data remains resident on a relational database server, that participates along with the ROLAP server in a true interactive three-tier architecture. "A ROLAP engine is a set of metadata-driven software components that generate SQL --statements and perform formulaic calculations based on users' multidimensional requests all at run time." [3]

ROLAP vendors feel that use of the relational model for multidimensional analysis is superior to use of a mutidimensional logical model. They point out that ROLAP products are much more open to standard relational OLTP architecture, and much more accessible to front-end application development, query, and reporting tools. And in a nutshell, they believe that ROLAP is just as capable as MDOLAP at providing high performance sophisticated analyses for users, while at the same time providing much

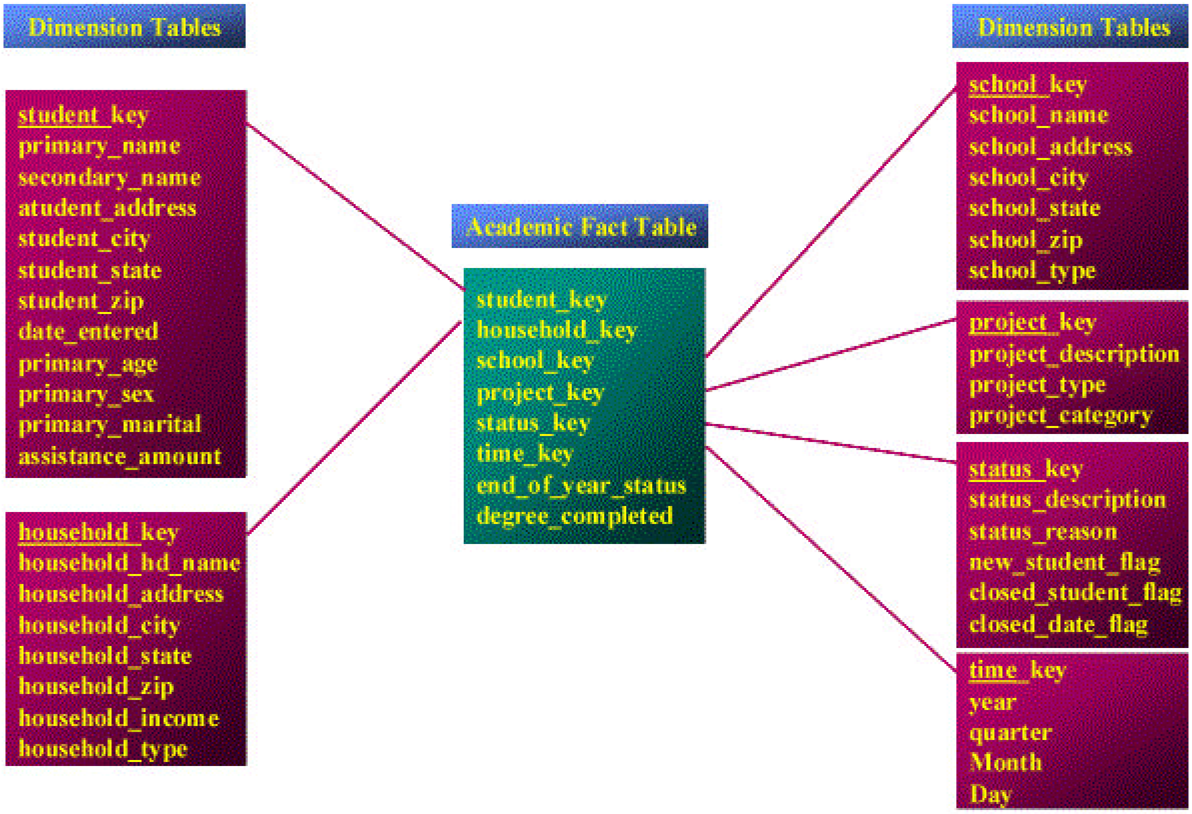


Figure Three -- A Dimensional Model Star Schema of A Student Academic Fact Database

greater scalability in relation to the number of dimensions and the amount of atomic data that enter an analysis.

Finally, ROLAP advocates believe that ROLAP is the only database technology that addresses the need for decision support in Very Large Database (VLDB) environments. In their view, ROLAP can provide unlimited scalability of dimensions and atomic data in the long run, because it is the only OLAP framework that can take advantage of parallel processing, bit-mapped indexing, and star join technologies, as well as further developments in the SQL paradigm.

VTOLAP --The Sybase IQ Alternative

Sybase IQ is based on different assumptions than either MDOLAP or ROLAP solutions to decision support systems or data warehouses. Since decision support queries often seek answers to general questions, they are oriented more to describing relations among attributes, fields, or columns of a database, rather than toward answering questions about individual or small groups of records. DSS processing is more column oriented than it is row or record oriented.

Sybase IQ begins with the assumption that such applications must be supported by data that is stored vertically, by column, rather than horizontally, by record. If data is stored vertically, rather than horizontally, it is possible to access a specific column named in a query directly, and since only the relevant attribute is involved, to retrieve data with a minimum of disk I/O. Sybase IQ combines such column-based physical organization with advanced compression techniques, flexibility in assigning I/O block sizes to queries, prejoin indexes, B-tree indexes, and advanced bit-mapped indexing that makes bit-maps useful even at high column cardinality values of 1,000 values or more, to produce a read only decision support engine that has been benchmarked at 10 to 100 times the query performance of conventional RDBMS.

Sybase IQ also has the advantage of small storage requirements. A data mart implemented in IQ may supplement an Oracle or Sybase back-end with a data mart requiring only 50% additonal space for the IQ transformed data mart. This is in marked contrast to ROLAP designs, which may take 50 or 100 times the space of the initial relational database on which they are based.

Even though Sybase IQ is a new release for Sybase, and is revised in its present form, it is not an entirely new product. A precursor to IQ called Expressway, was released in January of 1993 by Henco software, then an 18 year old software company that had specialized in document management and text retrieval. Expressway was well on its way to becoming quite successful, with notable applications at National Liberty and MCI, when Sybase acquired it in 1995. Expressway was submerged in Sybase for awhile, but after revisions and one hopes, considerable enhancement, was re-released by Sybase as IQ. Industry Press on IQ has been good both in its previous and present incarnations. [4]

The disadvantage of IQ is that it is based, like MDOLAP servers, on a nonrelational data structure, and is a proprietary product. On the other hand, IQ does interface with both Oracle and Cognos Impromptu, as well as other open relational products, while all ROLAP server-based engines that work along with relational servers, are also proprietary. As a practical matter, there may be no greater commitment to a proprietary product in selecting the VTOLAP IQ server than there is in selecting one of its MDOLAP or ROLAP competitors.

OLAP Evaluational Criteria

Most of the following evaluation criteria are meant to apply across product classes, though a few will be useful primarily for comparisons within a product class. The criteria are unweighted. To carry out a quantitative comparative evaluation, the criteria need to be structured further and to be prioritized through use of a formal method for setting priorities in the context of a specific application.

The method I like best for this kind of task is Saaty’s Analytic Hierarchy Process (AHP). you can learn about it and its myriad uses from the Expert Choice Internet site [5], and you may also find Zahedi’s application of AHP to the task of software evaluation [6] to be instructive in the present context.

- **Database size capacity of product relative to your data warehouse or data mart size requirements**

MDOLAP tools have been rapidly increasing their capacity to handle large databases as vendor competition has increased; but they generally have less capacity to handle sheer volume than ROLAP tools, or Sybase IQ. But the real question is: does the tool being evaluated have enough capacity for your DSS application?

- **Ability of tool to scale to the number of dimensions required by your DSS application.**

To evaluate this, you need to know what dimensions are required, and therefore to do a requirements analysis. In another White Paper, I developed a Systems Approach to Dimensional Data Modeling [7]. While that approach is specific to ROLAP development, the first three steps of the approach apply equally well to other types of OLAP development. The systems approach describes the steps to be followed in arriving at basic measurements tracked in OLAP applications, and dimensions that will be used to view and segment the units of analysis whose measurements are being tracked, into groups sharing values of attributes that are components of the dimensions. The approach will get you to the number of dimensions required for your application, and then you can compare tools according to their capacities to fulfill your requirements.

- **Ability of tool to support analyses against atomic data sets required by the DSS.**

Sometimes ad hoc analyses need to access the basic data underlying the data mart or data warehouse. Does the tool you are considering have this "drill-through" ability or not?

- **Ability of the OLAP tool to integrate directly with relational databases and non-numeric relational data.**

ROLAP tools have a built in advantage in this respect. But individual MDOLAP tools have established connectivity with various relational databases, as has Sybase IQ. You need to know how tools being considered interface with your relational database.

This criterion is a specification of the previous one applied to relational databases; Most OLAP tools will interface with Oracle, Sybase, and Informix, and many have no problem with MS SQL Server. But obviously care must be taken here to see that the OLAP tool supports the version of the database you prefer for the data warehouse.

- **Ability to perform calculations at run-time.**

This is important because the requirements of ad hoc queries are often unanticipated by designers and may not have been directly accomodated in the aggregation or compilation schemes of OLAP tools. The ability of a tool to perform calculations rapidly at run-time is a critical determinant of ad hoc query capability.

- **Data loading performance of the OLAP product.**

This criterion is more critical if the DSS is a frequently updated database. Depending on the business environment, loading can range from a daily affair on up to an annual update. Know what update frequency is required, and calculate the required carrying capacity of the tool accordingly. Match the requirement with benchmark results on the tool being considered.

- **Openness to standard reporting tools.**

In particular, can the product work along with the approved reporting tools in your organization? Or if you're not so constrained can it work with reporting tools you're considering?

- **Ad hoc query performance.**

Benchmarks need to be run to compare products on ad hoc query performance. Benchmarks should test query and reporting tools alone and in combination with OLAP products. Do OLAP products meet requirements for ad hoc query performance? How do they compare to one another when working along with your database?

- **Performance in running standard reports in conjunction with reporting tools.**

Again, benchmarks need to be run on comparative performance, and against any standard specified in the requirements analysis.

- **Training required for the OLAP product.**

Most OLAP products require only light end user training. But how do they compare to each other? And what kind of training is needed for power users and data warehouse/data mart administrative users?

- **Ability to produce full multiuser read-write applications with industrial strength security**.

This comes down to ability of the tool to prototype the DSS application. Is the required security available in the prototype? How does it benchmark?

- **Ability of the tool to integrate with your organization's enterprisewide environment by using standard middleware and client/server communications.**
- **Cost of ownership, training, and installation**.
- **Previous acceptance of the tool as an organizational standard or at least an option compatible with the enterprise computing environment.**

Absent such previous acceptance, selection of one tool over another may involve the often substantial bureacratic costs of a CIO or MIS-based approval process. This may greatly expand the lead time necessary to implement a data mart.

- **Ability to manage a three-tier decision support system in real-time.**

The OLAP tool needs to be able to manage query and report traffic simultaneously with communications with the data warehouse or data mart relational database server, with other analytical server-based data mining tools, with groupware applications, and with the internet.

- **Support in the tool for workflow automation.**

The OLAP tool may need to support a programmed workflow as an important component. Can it be integrated with workflow tools that produce this kind of application?

- **Support for tuning against tool produced performance statistics**.

Does the tool provide statistics that you can use for performance tuning? Without these you probably won’t be able to get the performance you need.

- **Extent of analytical capabilities relative to what your DSS application needs, and compared to other tools.**

OLAP tools vary widely in their analytical capabilities. Some produce basic analytical statistics, while others, provide substantial statistical analysis capability. What do you need in your application? What kinds of capabilities will you have outside of the OLAP tool? If these are substantial, how will the OLAP tool interface with your external analytical tools?

- **Support for OLE, and COM or CORBA distributed architecture.**

How distributed will your DSS architecture be? Tools will differ in the degree to which they can make use of distributed computing capabilities. If you’re planning to implement a distributed architecture, you’ll need an OLAP tool that supports one also.

- **Support for real-time query governing and flow control.**
- **Support for custom application development with standard front - end tools such as Visual Basic, Powerbuilder, and C++.**
- **Tool support for Internet deployment.**
- **Support for Dimension Table designs**

This criterion applies to ROLAP tools. It refers to support in constructing star and snowflake designs, and for queries against star or snowflake dimensional data models.

References

[1] The term OLAP is due to the father of the relational database movement, E. F. Codd. See his "On Line Analytical Processing," Arbor Software White Paper, 1993, for his 12 rules for OLAP.

[2] Microstrategy is already as large or larger than Logic Works, the manufacturer Of ERwin and BPwin.

[3] "OLAP: Scaling to the Masses," Information Advantage, Inc. White Paper, no date.

[4] See Steve Roti, "Riding High on Expressway 103" DBMS (July 1994), 90-92, and Richard Finkelstein, "Sybase IQ: Expressly for the Warehouse" Database Programming and Design (December, 1996), 47-49.

[5] http://www.expertchoice.com

[6] Fatemeh Zahedi, Intelligent Systems for Business: Expert systems with Neural Networks

(Belmont: CA, Wadsworth, 1993). Chapter 12.

[7] Joseph M. Firestone, "A Systems Approach to Dimensional Data Modeling," White Paper No.1, Executive Information Systems, Inc. Wilmington, DE, March 12, 1997 (available from the author).

Biography

Joseph M. Firestone is an independent Information Technology consultant working in the areas of Decision Support (especially Data Marts and Data Mining), Business Process Reengineering and Database Marketing. He formulated and is developing the idea of Market Systems Reengineering (MSR). In addition, he is developing an integrated data mining approach incorporating a fair comparison methodology for evaluating data mining results. You can e-mail Joe at eisai@home.com.

[Up] [Data Warehouses and Data Marts: New Definitions and New Conceptions]
[Is Data Staging Relational: A Comment]
[DKMA and The Data Warehouse Bus Architecture]
[The Corporate Information Factory or the Corporate Knowledge Factory]
[Architectural Evolution in Data Warehousing]
[Dimensional Modeling and E-R Modeling in the Data Warehouse]
[Dimensional Object Modeling] [Evaluating OLAP Alternatives]
[Data Mining and KDD: A Shifting Mosaic]
[Data Warehouses and Data Marts: A Dynamic View]
[A Systems Approach to Dimensional Modeling in Data Marts]
[Object-Oriented Data Warehousing]