A Hitchhikers Guide To OLAP

Paul Burton

Principal Consultant

Aurora Consulting Pty Ltd

Abstract

For many years, the quest for competitive advantages has prompted many organisations to attempt the paradigm shift from Online Data Processing (OLTP) to the exciting arena of Online Analytical Processing (OLAP). This paper attempts to demystify OLAP by explaining its key concepts and their applicability in today's business environment. The presentation concludes with a presentation of a live OLAP system.

About The Author

Since starting his IT career in 1993 Paul has been involved in a variety of projects and has gained in-depth experience in Data Warehouse and OLTP applications, and a wide range of technology products. Paul is a very highly skilled developer with extensive experience in relational database design, OLAP database design and many database and BI products. In recent years, Paul has played an integral role in a number of large Data Warehouse projects.

For more information on Paul or Aurora Consulting, visit http://www.aurora-consult.com.au or email info@aurora-consult.com.au.

Introduction

In spite of its phenomenal growth in recent years, there is an acute lack of literature in the subject of OLAP database design. The paper attempts to cover some key concepts of OLAP database design, thereby offering the readers a crash course on the subject.

What is OLAP ?

Firstly, we need to understand what OLAP is. Though closely related, OLAP should not be confused with the Data Warehouse or Data Mart.

A Data Warehouse is the repository that stores large amount of data extracted from a variety of source systems exist within the organisation. These can include sales systems, HR systems, Finance systems, etc. Depending on the requirements, data are stored in detail or summary form. Data Warehouse as a concept exists well before the term *Data Warehouse* is coined, and certainly well before OLAP as a concept existed. Some of terminology used in the earlier years for essentially similar concepts includes *Decision Support System* (DSS) and *Executive Information System* (EIS). Inmon, widely regarded as the "father of Data Warehousing", defined a Data Warehouse as "a subject-oriented, integrated, non-volatile, and time variant collection of data in support of management's decisions". The aim of the Data Warehouse is to bring together and integrate the information from the various source systems together so that they can be used to provide a combined and consolidated view of the organisation.

Online Analytical Processing (OLAP), on the other hand, is a technology by which Data Warehouse data are queried and analysed. It is no coincidence that Inmon's definition of a Data Warehouse is based on the content of the Data Warehouse and says nothing about the underlying implementation technology. A Data Warehouse does not necessarily need OLAP technology. In fact, in some circumstances, more traditional technologies such as the 3NF relational database can be an appropriate medium for a Data Warehouse.

OLAP products, also called Multidimensional products, derive their name from the multidimensional nature of most business enquiries. In a Research Note called "The Trouble with SQL", the Gartner Group illustrated the many dimensions in which business analysts typically examine data :

"Finance managers of large organizations don't ask, 'How much have we spent?' - a zero-dimensional question. Instead, they ask, 'How much have we spent on health benefits, by month, in division X, in each state, compared with plan?' - a five-dimensional question."

OLAP uses a multidimensional view of aggregate data to provide quick access to strategic information for analysis purposes. OLAP enables users to gain insight into data through fast, consistent, interactive access to different views of same information.

OLAP is thus a complementary concept to Data Warehouses. A Data Warehouse extracts, cleanses and stores the data, while OLAP presents Data Warehouse data as strategic information.

Who uses OLAP ?

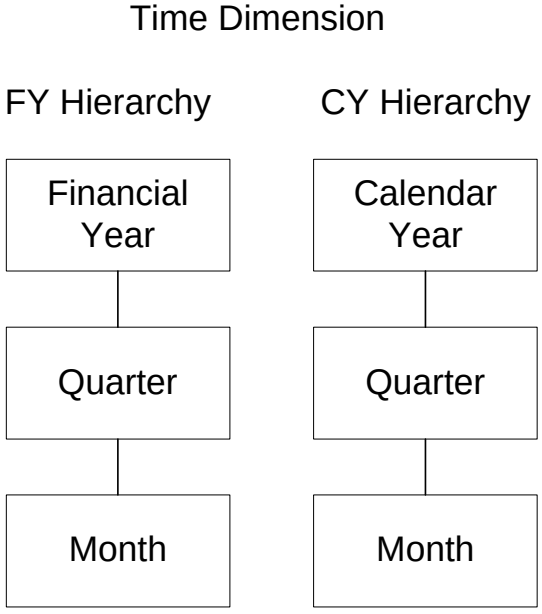
The multi-dimensi onal approach to analysis aligns the data content with the analyst's mental model, hence reducing confusion and lowering the incidence of erroneous interpretations. OLAP models also tend to concentrate purely on the organisations key performance indicators, reducing the amount of data "noise", and this making it easier to get to the heart of the analysis. It also eases navigation of the database, screening for a particular subset of data, asking for the data in a particular orientation and defining analytical calculations. Furthermore, because the data is physically stored in a multi- dimensional structure, the speed of these operations is many times faster and more consistent than is possible in other database structures. This combination of simplicity and speed is one of the key benefits of multi-dimensional analysis.

Once the power of OLAP is understood, its usefulness and impact within the organisation can be tremendous. The Finance Department may use OLAP for budgeting and financial modelling; the Sales Department may use OLAP for Sales analysis and forecasting; and the HR Department may use it for analysing overtime, sickness and leave liability.

OLAP Demystified

In designing and understanding an OLAP database, one inevitably encounters many OLAP terminology and concepts. Some the keys ones are :

- **Measure or Fact** A measure (or fact) is a data variable that stores business information, classified by one or more mutually independent categorising structures. Eg. Budgeted Expenditure, Actual Expenditure, number of widgets sold.
- **Dimension.** A dimension is a categorising structure by which business measures are classified for analysis purposes. Eg. Time, Product, Geography and Gender
- **Dimension Value.** A dimension consists of one or more dimension values, each representing a distinct category of classification. For example, in a Gender dimension, there may be 2 dimension values - Male and Female.
- **Dimensionality.** When a measure is classified by a number of dimensions, we say that the measure is *dimensioned* by those dimensions. The dimensionality of the measure is the combination of the dimensions that classified that measure. As a general rule, the dimensionality of a measure should not exceed 7. Otherwise, the resultant measure is likely to be too complex for comprehensive analysis.
- **Hierarchy.** A hierarchy is a parentage structure by which dimension values are organised. A dimension can be organised into one or more hierarchies. For example, a Time dimension may have 2 hierarchies - Financial Year and Calendar Year.



- **Level.** A level is a logical grouping of dimension values sharing the same level of abstraction within a hierarchy. Each level represents a level of aggregation within the dimension. In the earlier example of the Time dimension, the FY hierarchy consists of the 3 levels of Financial Year, Quarter and Month; while the CY hierarchy has the 3 levels of Calendar Year, Quarter and Month. As a general rule, the number of distinct dimension values at a particular level within a hierarchy should be restricted to 20 or less. Otherwise, one should consider adding an additional level to group related dimension values into higher level classifications.
- **Parent and Child.** Once grouped into levels and hierarchies, dimension values, dimension values are related to each other as parents, children or siblings. The parent is the dimension value that is one level up in a hierarchy from another dimensional value. The parent value is a consolidation of all of its children's values. Children are dimension values at a particular level in the hierarchy, one level down from another dimension value. Children can themselves be parents and can also have more then one parent leading to complex multi-hierarchical dimensions. For example, in the Time dimension, Q1 2002 is :
  - The parent of January 2002, February 2002 and March 2002
  - A child of 2001/2002 on the FY hierarchy
  - A child 2002 on the CY hierarchy
- **Cell.** A cell is a single data-point that occurs at the intersection defined by selecting one dimension value from each dimension in the measure's dimensionality. For example, if sales is dimensioned by Time and Geography, the cell intersected by February 2002 and Perth represents the sales made in Perth in February2002.

	Jan 2002	Feb 2002	Mar 2002
Sydney			
Perth		$21,943	
Adelaide			

- **Cube.** A cube is the collective term for the multi-dimensional array of data cells.

- **Drill Up, Drill Down**. Drilling up or down is a specific analytical technique whereby the user navigates among dimension levels, from the most summarized (up) to the most detailed (down). The drilling paths are defined by the hierarchies within the dimensions. The following diagram illustrates an example :

	Jan 2002	Feb 2002	Mar 2002
Sydney	$33,298	$30,243	$31,213
Perth	$18,921	$21,943	$23,181
Adelaide	$15,216	$12,134	$10,341

Drill up to Quarters

	Q1 2002	Q2 2002
Sydney	$94,754	$80,223
Perth	$64,045	$35,912
Adelaide	$37,691	$35,371

Drill down to Perth Suburbs

	Q1 2002	Q2 2002
Morley	$24,704	$20,244
Como	$19,040	$15,060
Manning	$20,301	$ 608

- **Drill through**. OLAP analysis can be used to quickly identify anomolies or trends in the summary level data within a cube. However, there is often a requirement to dig deeper for more detailed information that are not in the cube. This can be facilitated by a drill through operation from the cube into relational Data Warehouse data from which the cube is based. The following is an example of a drill-through report :

Drill through to Manning sales data for Q2 2002

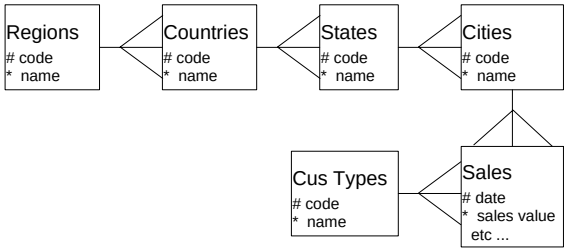
Date	Sales
1 Apr 2002	$194
2 Apr 2002	$202
3 Apr 2002	$212
4 Apr 2002	$ 0
5 Apr 2002	$ 0
6 Apr 2002	$ 0
7 Apr 2002	$ 0
…	
Total	$608

(It turns out that the Manning branch was burnt to the ground on 4 Apr 2002.)

Putting It All To Together

Perhaps the best way to illustrates these OLAP terminology and concepts is by an example :

A fast food chain hires a consultant to develop a Business Intelligence system to provide its senior management with timely and accurate information for making strategic business decisions. One of the most important information required for strategic decision-making is information relating to the customer base. Below are some of the tables used by the Sales System :



From the diagram above, the consultant extracts the following measures :

- Sales Value
- No of Sales

The consultant then determines that these measures can be classified by the following dimensions :

- Types of customers (i.e. drive-through or counter sales)
- Geography of the outlets (i.e. the city)
- Value range of sales (i.e. <$5, $5 to $10, $11 to $20, >$20, etc.)
- Date of sales (i.e. months, quarter, year)

Moreover, the Geography dimension can be organised as a 4-level *hierarchy* : Region-Country-State-City. Once structured in a multi-dimensional manner, the measure can be examined across its multiple dimensions simultaneously. For example, one can analyse on whether there is any visible correlation between the types of customers and the location of the outlets; or observe the spread of the value range of sales across the various cities.

Conclusion

Business Intelligence (BI) applications are a significant driving force in many organizations. They provide the ability to gain insight into the

organization's core business and internal operations, and allow the organization to react quickly to a changing environment. It is important to note that while OLAP has a vital role to play, it is but a part of a bigger BI puzzle. OLAP should be seen as a part of the overall knowledge management strategy that incorporates a Data Warehouse, one or more Data Marts, and well-documented business rules (on how the data are to be extracted, validated and interpreted). Often, there is a vast amount of work involved in identifying what is to be counted (the measures), how to classified (the dimensions) and how the data is to be extracted, transformed and stored (Extraction, Transformation and Loading rules). Only when this understanding is reached can meaningful analysis begin in earnest.

References

Inmon, W H. *Building The Data Warehouse.* Second Edition. John Wiley & Sons 1996.