CS6905 Part II: OLAP and Data Warehousing

OWEN KASER, Dept. CSAS, UNBSJ

DANIEL LEMIRE, UQAM

These slides are available on the web.

http://sjwebserver.unbsj.ca/~owen/courses/OLAP-2006/

Introduction to OLAP

Sources: Daniel's intro slide from 2003; Owen's presentation to NRC, 2003.

Also: [Pen02, ola]

Overview

- ✔ Review of the industry
- ✔ Motivation through example
- ✔ Definitions!!!

OLAP is important?

15.7% growth in 2004 (most recent result)

Source: OLAP Report of 28 June 2005. (http://www.olapreport.com)

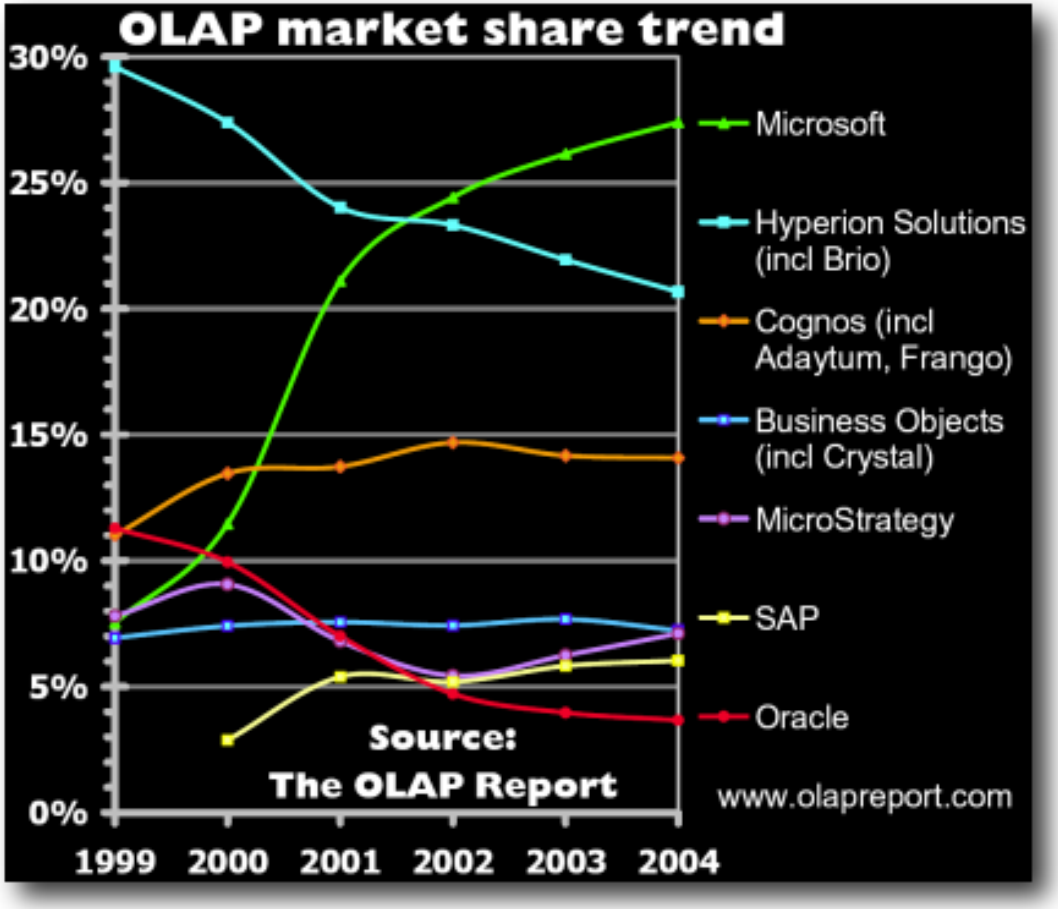
Historical Perspective

1970	Codd proposes relational model
$\gtrsim$ 1980	SQL becomes a commercial success (Oracle, IBM)
1993	Codd coined OLAP, Excel offers *Pivot Tables*
1997	MOLAP vs ROLAP debate
1999	SQL-99 offers some OLAP functionality

Industry standards

Name	Status	Platform	Proponent
OLE DB	In use	Wintel	Microsoft
XML Analysis	seems stalled?	SOAP	Microsoft, Hyperion
JOLAP	never approved?	Java (J2EE)	IBM, Oracle, Hyperion, Sun

Who sells OLAP?



Open-source OLAP: Mondrian

Mondrian (http://mondrian.sourceforge.net) uses one of several relational databases for data storage. Written in Java.

It supports MDX (OLE DB) query language. It has a JOLAP-ish API. It has an XMLA interface.

Mondrian installation is *not* trivial. My experience: the implementation may be incomplete or buggy.

Nevertheless, we'll use it.

OLAP: Sales Example

Assume one (real-valued) measure value, `Sales Amount` and dimensions `Item`, `Place`, `Month`.

Example fact: “Iced Tea was sold in Auckland in January”. Measure: $20k

Maybe *no* fact about “Iced Tea in Auckland in August”.

Mapping $\texttt{Item} \times \texttt{Place} \times \texttt{Month} \longrightarrow \texttt{Sales Amount}$ is our cube.

Dimensions

Dimension values commonly organized into *containment* hierarchies.

Example:

- Schema: `Place` is organized by `City`, `Province`, `Country`, $\top$.
- Instance: Invercargill is in Otago, Christchurch is in Canterbury, Boston is in Massachusetts, Otago and Canterbury are in New Zealand, all places are in $\top$ ...

Example OLAP session

Show total sales of `Iced Tea`. "Dice" by item Iced Tea.

Answer "$32M". Seems low. Who's not drinking?

Next query asks for total sales of `Iced Tea` by `Country`. "drill down" on `Place`

Data returned in a 1-d table.

Australia	Canada	New Zealand	. . .	USA
5	2	1		20

What? Only $1M for New Zealand??

OLAP session continues

OK, show me Iced Tea sales for New Zealand broken down by month. further refining the dice, drilling down on month

Data returned in a 1-d table.

Jan	Feb	Mar	Apr	May	Jun	Jul	. . .	Dec
0.2	0.1	0.1	0.05	0.05	0.03	0.02		0.15

Weird, sales dropping in June. Maybe it's geographic.

OLAP session continues

drill down some more

Show me Iced Tea sales for New Zealand cities, by month.

Big Query Answer

Data returned in a 2-d table (in $k),

	Jan	Feb	Mar	Apr	May	Jun	Jul	. . .	Dec
Auckland	20	20	20	20	20	18	18		20
Blenheim	5	5	4	4	4	4	3		5
Christchurch	10	10	10	9	8	8	8		11
. . .									
omitting 30									
cities									
. . .									
Wellington	2	2	2	1	1	1	1		2

Lots of variability, and drowning in data, so roll up on Place

OLAP session continues

Show me Iced Tea sales for N.Z. “provinces”, by month

Data returned in 9-row 2-d table (in $k),

	Jan	Feb	Mar	Apr	May	Jun	Jul	. . .	Dec
Auckland	29	29	29	29	29	28	27		30
Canterbury	25	25	24	18	14	12	12		25
. . .									
Otago	30	30	25	15	5	0	0		27
. . .									

Aha: Not many sales in Otago in June and July. Guess it’s too cold then.

OLAP session concludes

Decide to run the “10 unconventional uses for Iced Tea Mix” ad in Otago.

But what is OLAP exactly?

Short answer: a marketing term more *catchy* than **analysis using a multidimensional database**.

Providing OLAP (On-Line Analytical Processing) to User-Analysts: An IT Mandate. 1993

Old URL: `www.essbase.com/whitepaper/olap/olap.pdf`

See also [Pen02].

Some of Codd's defining conditions

- ✔ Multidimensional Conceptual View
- ✔ Generic Dimensionality
- ✔ Unlimited Dimensions and Aggregation Levels

Some of Codd's defining conditions

- ✔ Unrestricted Cross-Dimensional Operations
- ✔ Consistent Reporting Performance
- ✔ Dynamic Sparse Matrix Handling

6 more rules (total 18) added later.

Fast Analysis of Shared Multidimensional Information

- ✔ Codd's list is long and tedious
- ✔ 12+6 rules defining what OLAP is lot to re-member
- ✔ Nigel Pendse proposed his FASMI definition in 1995, has been widely adopted [Pen02]

The FASMI "definition"

- ✔ FAST: responses to users within 5s
- ✔ ANALYSIS: can cope with any business logic and statistical analysis
- ✔ SHARED: security requirements and concurrent update locking
- ✔ MULTIDIMENSIONAL: multidim. conceptual view of data
- ✔ INFORMATION: all data/derived information needed

Decision Oriented Concepts

Data Mining : pattern discovery $\rightsquigarrow$ rules, models $\rightsquigarrow$ rule-based decision

$\triangleright$ Examples: regressions, neural nets, decision trees, clustering

OLAP: analysis $\rightsquigarrow$ descriptive knowledge (human) $\rightsquigarrow$ decision

Definitions

Variable A unit-bearing data type, either measured or derived.

Attribute Information associated with an object.

Dimension Collection of objects of the same type.

For our purposes, Variable = Attribute.

Dimension versus Variable

	weight	height
John	160lbs	1.8m
Maggy	125lbs	1.4m

Definitions

To Aggregate The process of combining two or more data items into a single item.

Measure A unit-bearing data type (and how to aggregate it).

Cell A measure associated with one and only one member from each of multiple dimensions.

Hypercube or Data Cube A multi-dimensional schema formed from the cross-product of a number of dimensions.

References

[ola] The OLAP report (website). online at http://www.olapreport.com/. checked February 20, 2006.

[Pen02] Nigel Pendse. What is OLAP? online at http://www.olapreport.com/fasmi/, 2002. checked February 20, 2006.