



PERCONA
Performance Consulting Experts

Introduction to OLAP

Presented by:

Justin Swanhart,
Percona

What is BI

- Business Intelligence is about providing insight about business operations
- Combination of software technologies
- Covers multiple areas
 - Data warehousing
 - Data mining
 - Reporting tools
 - OLAP Analysis

Data warehousing

- What is a data warehouse?
 - Archived information from an operational system such as an ERP system or an ad serving platform
 - Usually time-invariant, that is, data is added to the system but it is not deleted or updated
 - Often records information from 'systems of record'
 - The data warehouse reflects the real world as closely as possible.

Data Warehouse II

- Highly normalized schema is optimized for insertion of data
- Difficult to write queries on without intimate knowledge of the data set

Data Marts

- The data used for OLAP analysis is almost always stored in a star schema
- A star schema built from operational data for the (near) exclusive use of OLAP applications is called a 'data mart'
- Often a department will maintain or operate a data mart

Data Mining

- Vast amounts of data are collected
- Patterns are identified
- Example
 - Market basket analysis
 - Physical and experimental models
 - Fraud detection

Reporting

- Usually used against a data warehouse or operational schema
- Reports and breakdowns
 - Sales commission reports
 - Business objectives and performance analysis
- Usually focused on presentation and formatting data for printing or reporting to end users

OLAP

OnLine Analytical Processing

- Multi-dimensional analysis
- Specialized tools
 - ROLAP/MOLAP/HOLAP
 - Reporting tools

Multi-dimensional analysis

- Analysis is always done based on a single subject matter, such as sales
- Data is broken down into *measures* and *dimensions*.
- Data is usually presented in 'pivot' format, summarized over two axis.

Sales Analysis

Dimensions

- Product
- Customer
- Date

Measures

- Sale Amount

Product	Customer	Date	Sale_Amt
Apple	FoodPlace	1/1/2010	100.56
Apple	FoodPlace	1/2/2010	100.56
Orange	O'Mart	1/3/2010	10.99
Bannana	Scuttles	1/3/2010	425.00

OLAP Models

- Multi-dimensional is usually stored in a special purpose database or schema
 - ROLAP
 - MOLAP
 - HOLAP (hybrid)

ROLAP

- Stores data in an RDBMS
 - Oracle and Oracle Exadata
 - Mysql
 - Infobright ICE
 - Calpoint InfiniDB community edition
 - SQL Server
- Stores information in a 'star' schema
- Can scale to very large data sizes
- Frequent updates are possible

MOLAP

- Data is stored in a specialized multidimensional store
 - Many MOLAP databases are designed to retain the entire data set in memory. It may not be possible to manipulate data sets larger than memory
 - Some MOLAP databases are client based and feature significant in-memory compression for performance
- Most (or all) query answers are precomputed from the data store for very fast end-user response

MOLAP - cont

- This pre-calculation is expensive and makes loading slower. This is called 'processing time'.
- Since loading data and processing are expensive, real-time loading of data is not common in MOLAP platforms.

Two Paradigms

- Bottom up approach builds normalized data warehouses from departmental datamarts and other sources
- Top down approach builds data marts from normalized data warehouses. This allows marts to easily be rebuilt, but requires a lot of planning and database infrastructure.

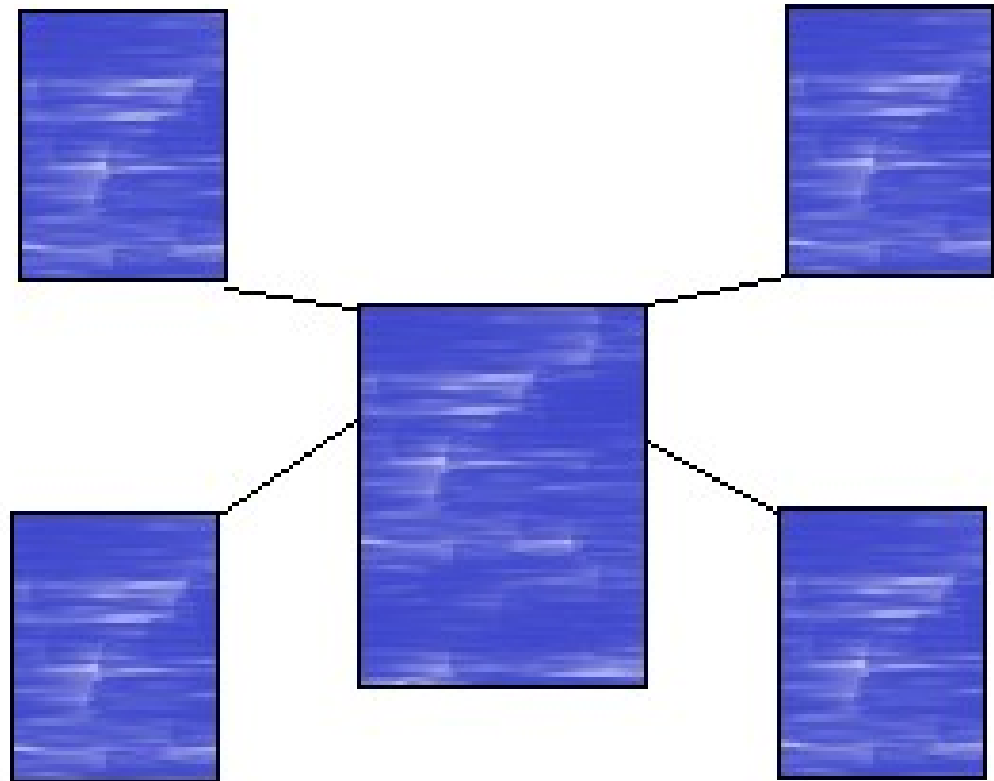
Sales Analysis — Star schema

Dimensions

- Product
- Customer
- Date

Measures

- Sale total



Fact table

- The fact table contains the measures, and columns linking each row to a single row in each dimension table. This is also known as a foreign key.

product_id	INT
customer_id	INT
sale_date	DATE
sale_total	DECIMAL(5,2)

Dimension tables

- The dimension tables contain information (usually textual information) about the dimensions

product_id	INT
name	varchar(20)
category	varchar(10)
weight	float

Slowly Changing Dimensions

- The dimension table contains start/end times for each row, capturing historical information

product_id	INT
start_date	date
end_date	date
name	varchar(20)
category	varchar(10)
weight	float

Hierarchies and Levels

- Hierarchies are used for 'drilling'
- Hierarchies consist of Levels
- Dimension tables may contain multiple hierarchies
 - Most common example is the DATE dimension
 - YEAR → MONTH → DAY
 - YEAR → QUARTER → WEEK → DAY
- Data is aggregated to a particular Level of a particular hierarchy.
- All dimensions have an 'All' level, which is usually the default aggregation level for the dimension.

Measures

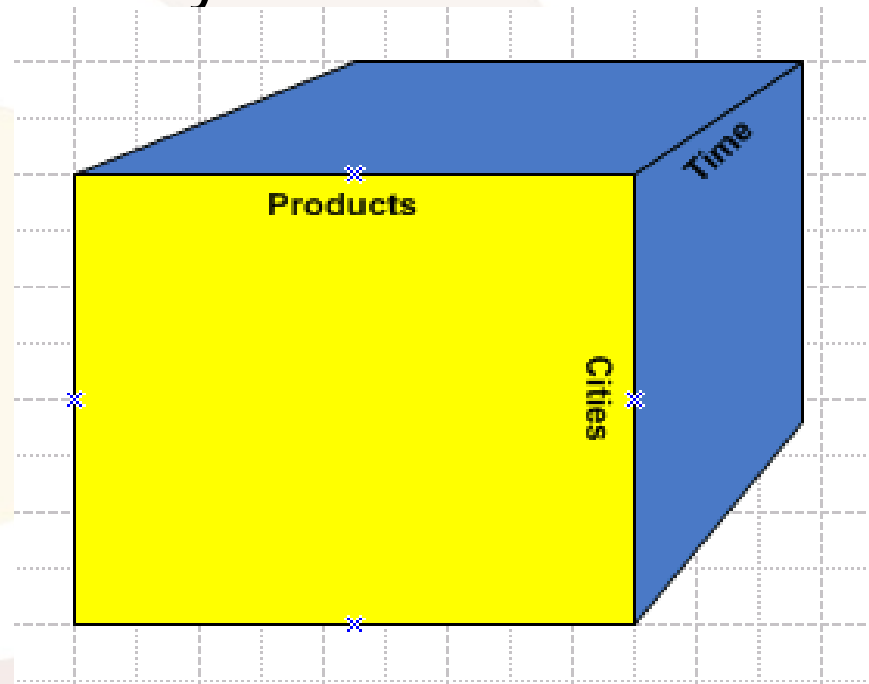
- Measures are always aggregated
 - Default aggregation is SUM.
 - MIN/MAX/AVG/COUNT and COUNT(DISTINCT) may be used
- Measures are stored in the fact table, and only in the fact table.

Measures - cont

- Measures may be aggregated to a default 'grain' in the fact table. For example, all impressions/clicks for a particular advertisement may be aggregated to the day level before being inserted into the fact table.
- Measures can be 'calculated measures'. These measures are built from expressions.

OLAP Cube

- The result of a join and the dimension tables produces a three dimensional cube
- The aggregated measures at the 'intersection' of the dimensions is analyzed.



Cross tabulation

- Used to visualize data from the OLAP cube
- Very common in spreadsheets (pivot table or pivot report)
- While it can work with any number of dimensions, it displays data in only two dimensions (or axis)

Cross tabs - cont

	A	B	C	D	E	F	G
1	Region	Gender	Style	Ship Date	Units	Price	Cost
2	East	Boy	Tee	1/31/2005	12	11.04	10.42
3	East	Boy	Golf	1/31/2005	12	13	12.6
4	East	Boy	Fancy	1/31/2005	12	11.96	11.74
5	East	Girl	Tee	1/31/2005	10	11.27	10.56
6	East	Girl	Golf	1/31/2005	10	12.12	11.95
7	East	Girl	Fancy	1/31/2005	10	13.74	13.33
8	West	Boy	Tee	1/31/2005	11	11.44	10.94
9	West	Boy	Golf	1/31/2005	11	12.63	11.73
10	West	Boy	Fancy	1/31/2005	11	12.06	11.51
11	West	Girl	Tee	1/31/2005	15	13.42	13.29
12	West	Girl	Golf	1/31/2005	15	11.48	10.67

Cross tabs - cont

Sum of Units	Ship Date ▼					
Region ▼	1/31/2005	2/28/2005	3/31/2005	4/30/2005	5/31/2005	6/30/2005
East	66	80	102	116	127	125
North	96	117	138	151	154	156
South	123	141	157	178	191	202
West	78	97	117	136	150	157
(blank)						
Grand Total	363	435	514	581	622	640

MDX

- MDX is a SQL-like language that is used to query OLAP cubes.
- Looks like SQL, but is actually quite different, sometimes even confusing
 - Three main axis: ROWS, COLUMNS, WHERE
 - A dimension may only appear in ONE axis!
- Identifiers should be enclosed in square brackets
- Main clauses:
 - SELECT, ON COLUMNS, ON ROWS, FROM, WHERE

MDX Example

```
SELECT  
[Product].[Categories].Members ON COLUMNS,  
FILTER (  
    [Customers].Members,  
    [Sale_total]>5000  
) ON ROWS  
FROM [sales_cube]  
WHERE [Year].[2009]
```

Slicing

- When you apply a **WHERE** clause, you take a 'slice' of the cube, only displaying a small subset of the data.
- Remember that any dimension on which you slice can not be placed on the **ROWS** or **COLUMNS** axis.
- You do not slice on particular expressions, but on values, or **SETs** of values:
 - `WHERE [Year].2002`
 - `WHERE {[Year].2002, [Year].2003}`
- **SETs** are enclosed in curly braces

Leverage hierarchies

- A hierarchy member can have children
- You can use this to your advantage if you want to limit display of certain ROWS or COLUMNS

SELECT

[Customers.Region].[NE].CHILDREN ON COLUMNS,
[Products].Categories.Members ON ROWS

FROM sales_cube

WHERE [Year].[2008]

Dicing

- Refers to rolling a dice (a cube) to change the face
- Switching from sales by customer to sales by salesrep, or from sales by region to sales by product category are examples of dicing.
- A dice may define a sub-cube of the original cube
- Can be combined with slicing

Drilling

- Drilling changes the hierarchical level to which the data is aggregated
- If looking at sales by date, with a hierarchy of 'Year → Quarter → Month → Day' would start by default at the Year level, but you could 'Drill down' to see aggregation at each level, or drill up to a higher level
- The special 'drill through' method applies only to measures and displays the unaggregated fact values for the measure.

Mondrian

- ROLAP server
- Java based
- Created by Pentaho, and released as free open source
- Communicates via major analytics interfaces
- Uses JDBC to communicate with the RDBMS
- Collection of fact tables and dimension tables that form the cube is maintained in a Mondrian 'Schema' file, structured as XML.

Aggregate Tables

- Aggregate tables store pre-summarized information in the database
- Allows a ROLAP system to perform closer to MOLAP by precomputing data
- Often expensive and difficult to maintain
- Creation is aided by Pentaho Aggregate Designer
- Mondrian can use the aggregate tables as long as they have been registered in the schema.
- Aggregate tables are excellent candidates for materialized views

QUESTIONS

More Info

- <http://www.mysqlperformanceblog.com/2010/07/12/intro-to-olap/>
- <http://en.wikipedia.org/wiki/Olap>
- Pentaho Solutions — Roland Bouman, Jos van Dongen, Wiley press ISBN: 978-0-470-48432-6
- Building the Data Warehouse, fourth ed. – William H. Inmon, Wiley press ISBN: 978-0-7645-9944-6
- Flexviews, Materialized views for MySQL - <http://flexviews.sourceforge.net>