Next-Generation OLAP
The Future of Dimensional Analysis

John O’Brien
Zukeran Technologies
Executive Summit - August 2008
TDWI Colorado Chapter – October 2008
Agenda

• OLAP Overview

• A brief history of OLAP vendors and products

• The next-generation technologies and styles

• What does the future hold
OLAP OVERVIEW
Online Analytical Processing

Ability to quickly answer analytical queries that are multi-dimensional in nature.

1. Interactive experience requirement
   - A few seconds for every answer

2. “Framed” business questions
   - How much of what happened when by whom?

3. “Slice and Dice” means Pivot tables at click speed
   - Microsoft Excel with plug-ins
Fundamental principles

• Dimensional:
  – Logically modeled star schemas with Facts and Dimensions

• Aggregated:
  – Performance from pre-calculating aggregates
    • Sum, Average, Min, Max, Count, Distinct Count
  – View selections to minimize over calculating every combination

• Hierarchical:
  – Levels of aggregation per dimension
    • Example: day -> week -> month -> quarter -> year
  – Multiple Hierarchies on the same dimension
    • Example: Fiscal and Calendar year, Sales regions and geography
OLAP in Mixed BI Workloads

- **END-USERS**
  - Executives, Managers, Staff (70%)
  - External Users (10%)

- **AUTHORS**
  - Power Users (15%)
  - Dashboards/Scorecards
  - Power Users
  - Dimension analysis (OLAP) Forecasts & Plans
  - Cluster Models Predictive Models

- **IT Developers**
  - Pixel Perfect Reports Static Reports
  - Parameterized Reports Interactive Reports

- Analysis (Depth)
  - Interactivity, Performance, Complexity

- Reporting (Breadth)
  - User and Data Scalability

- **TDWI Executive Summit**
OLAP Taxonomy

Mainstream:
• MOLAP – Classic multi-dimensional cubes
• ROLAP – Relational database used
• HOLAP – Hybrid combination of MOLAP/ROLAP

Minority Variations:
• DOLAP – Desktop OLAP
• WOLAP – Web-based OLAP
• RTOLAP – Real-time OLAP
MOLAP Cubes

• Multi-dimensional database server optimized for;
  – *Highest performance from specialized indexing and storage optimizations*
  – Aggregating data
  – Storing and compressing data
  – High user concurrency

• Trade-offs:
  – *Latency: loading and aggregating takes time*
  – *Scalability: combinatorial limits and size keep cubes small*
  – *Typically read-only database*
Relational OLAP

• RDBMS optimized for DW workload
  – Star-schema or Snowflake data models
  – Leverages summary tables and materialized views
  – Closer to real-time data loads
  – No additional database required
  – Scales as high as RDBMS can

• Trade-offs:
  – Slower performance since storage is not optimized
  – Higher data management for summary tables/mat’l views
  – Limited to SQL syntax
Looking back to see the future

A BRIEF HISTORY OF OLAP
45 years and here to stay

Some of the highlights:
1970 – First OLAP product Express introduced by Information Resources
1982 – Comshare System W launched first OLAP for financials
1984 – Metaphor launched first ROLAP
1990 – Cognos PowerPlay launched
1992 – Essbase launched (“Extended Spread Sheet dataBASE”)
1993 – Codd white paper coined the “OLAP” term
1994 – MicroStrategy DSS Agent launched (multi-pass SQL)
1995 – Oracle acquires Express
1996 – Business Objects 4.0 launches DOLAP
1998 – IBM DB2 OLAP Server released
1998 – Hyperion Solutions formed from Arbor and Hyperion merger
1999 – Microsoft OLAP Services shipped (Analysis Services in 2000)
2000 – XML for Analysis announced
OLAP history continued

2002 – Oracle ships integrated OLAP server (integration of Express)
   (but not usable until 2006, 11 years later after Express purchase)
2003 – Consolidation starts
   - Business Objects buys Crystal Decisions
   - Hyperion Solutions buys Brio Software
   - Cognos buys Adaytum
   - GEAC buys Comshare
2004 – Excel add-ins go mainstream
2005 – Pentaho buys Mondrian (Open source BI environment)
2006 – Palo launched

2007 – Oracle buys Hyperion (just when it delivers 11g embedded OLAP)
2008 – IBM buys Cognos
2008 – SAP buys Business Objects
2006 OLAP Market Share

OLAP market share trend

Source: The OLAP Report

Worldwide total OLAP market size

Source: The OLAP Report

www.olapreport.com
OLAP API wars

• XML for Analysis (XMLA) - 2000
  – *Becoming the most widely accepted of API standard*
  – *Microsoft led and evolved from OLE DB for OLAP*
  – *Not tied to a platform, support web and server to server calls*
  – *Essbase and Oracle OLAP are now leveraging MDX*
  – *Native API in SQL Server Analysis Server 2005*

• JOLAP – 2000
  – *Java based OLAP API from multiple vendors, led by Hyperion*
  – *Quietly forgotten as not implemented well*

• OLAP Council and Analytical Solutions Forum are now forgotten
Historical OLAP Challenges

• MOLAP Deficiencies:
  – Limited scalability leads to cube farms
  – Expertise required to implement
  – Sparse data causes problems
  – Build times are slow

• ROLAP Deficiencies:
  – Slow query performance
  – High cost of aggregate data management
  – Materialized views take time to implement for reports

• DOLAP Deficiencies:
  – Desktops can lead to inconsistent reports with data extracts
Traditional MOLAP Matrix weakness

- *Data explosion from matrix sparsity*
- *Memory limitation of addressable matrix values*
- *Update/Merge very slow*

Some location may have no data while others have majority

All Products, Customers, Channels, Promotions, etc. do not sell equally, or at all

Metrics (e.g. Sales) can fluctuate for day of week, month, seasonal

Metric1: Sales
Metric2: Cost
Metric3: M1-M2
...
Metric100

Time

Geography
Partitioning cubes to make it work

• Partitioning led to “cube farms” and data management nightmares similar to materialized views

• Horizontal Partitioning:
  – *Store only summary data in cubes and delete detail data*
  – *Detail data is lost for analysis purposes*

• Vertical Partitioning:
  – *Build cubes for federation*
  – *Cubes subdivided by a dimensional group(s)*
  – *Example: cube for east/central/west and retail/wholesale per month (144 cubes per year without grand totals)*
  – *Updating data across cubes is a nightmare*
  – *Prone to data duplication in multiples cubes without detection*
The next 3-5 years

NEXT GENERATION OLAP
The Next-Generation (here today)

- BI and DW Appliances
- Columnar Databases
- BI tools with in-memory data
- Non-matrix technologies
- Self-service BI methods
- Hybrid OLAP goes mainstream
Microsoft BI Summit: Oct 6, 2008

• Project Madison 2010 (18-24 months):
  – SQL Server + DATAllegro
  – hardware will come from "the usual, industry-standard hardware partners, including HP, Unisis, Dell and Bull"

• Project Gemini 2010:
  – another storage mode for Microsoft SQL Server Analysis Services with access via MDX
  – Tapping into in-memory capabilities built into Analysis Services with the aid of an add-in to the Excel client, users will reportedly be able to slice, dice and filter vast data sets (into the millions of rows) without aggregations or prebuilt cubes developed by IT
Oracle Open World Announcements

• Oracle Database Machine – Exadata
  – **Processing Nodes and Storage Nodes**
  – **Data is stored in block at the storage nodes**
  – **Every storage node is accessible from each processing node**
  – **Row selection and column projection performed at the storage node gains parallel performance and reduce I/O to processing nodes in a RAC configuration**
  – **Exadata and NeoView both from HP?**
  – **Processing node: 8 DL360 G5 servers**
  – **Storage node: 14 HP ProLiant DL180 G5 servers**
  – **Sold separately: Exadata s/w, Ent Edintion w/ RAC & Partitioning**
  – **Pre-configured RAC?**
  – **Support: Oracle, HP, Linux?**
Data Warehouse Appliances (DWA)

• Relational OLAP powered by MPP architecture
  – Massively Parallel Processing (MPP) database
  – Excels at very large database set operations found in dimensional models
  – Ease of use “Appliance” form factor
  – Easily handles 100TB

• For more information
  – See TDWI Marketplace: Solution Gateway: DW Appliances
  – See B-Eye-Network Channel: DW Appliances and Architectures
DW Appliance Architecture

Massive Parallel Processing of SQL is used to gain performance

Appliance “in a Node” form factor

BI Tool
Or
End User Query

Appliance “in a box” form factor

Table

Disk/CPU/Memory/Network/SQL

MPP Control

...
# DW Appliance vendors

## Data Warehouse Appliances

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Product</th>
<th>Differentiators</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATAlegra</td>
<td>DATALegro V3</td>
<td>Ingres-based database prepackaged with EMC, Dell and (in Europe) Bull hardware. Also runs on other commodity hardware.</td>
</tr>
<tr>
<td>Dataupia</td>
<td>Satori Server</td>
<td>Software/hardware combination natively supports apps running on Oracle, MS SQL Server and IBM DB2. Sold in 2 TB blades.</td>
</tr>
<tr>
<td>Greenplum</td>
<td>Database G3</td>
<td>Software-only DB based on Bizgres/PostgreSQL. Hardware partners include EMC, NetApp, Dell, HP and Sun.</td>
</tr>
<tr>
<td>HP</td>
<td>NeoView</td>
<td>Complete software/hardware system (HP Nonstop DB/HP servers and storage). Geared to operational BI (mixed queries/high volumes of users/simultaneous loading and querying).</td>
</tr>
<tr>
<td>IBM</td>
<td>IBM Balanced Warehouse</td>
<td>Available in C, D and E configurations geared to mid-size, enterprise and large enterprise deployments, respectively. Combines IBM databases and hardware.</td>
</tr>
<tr>
<td>Kognitio</td>
<td>WX2</td>
<td>UK-based, eight-year-old category pioneer now entering US market. Runs on commodity hardware.</td>
</tr>
<tr>
<td>Netezza</td>
<td>Netezza Performance Server</td>
<td>Software/proprietary, query-optimized hardware. Early marketshare leader with 125-plus customers.</td>
</tr>
<tr>
<td>Sun</td>
<td>The Data Warehouse Appliance</td>
<td>Greenplum database prepackaged with Sun servers and storage.</td>
</tr>
</tbody>
</table>

*Oracle appliance-style configurations not listed.*
Columnar databases

- Storing data in columns rather than rows allows the database to focus on what’s in the query
- Compression is good due to consistent repeating values in the same column
- Returning less data from columns reduces I/O
- Similar to working with indices

Good for OLAP? – Discussion…

- Aggregates – yes – “what’s the sum, avg of sales for last month?”
- Referenced – maybe? – “show me all customers addresses with sales over $100 of sporting goods in small cities?”
- Loading – maybe – how is load performance? Full reloads only?
# New Columnar Database Vendors

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Product</th>
<th>Differentiators</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Calpont</strong></td>
<td>CNX Data Warehouse Platform</td>
<td>Compatible with existing Oracle database environments. Modular architecture for incremental scaling.</td>
</tr>
<tr>
<td><strong>Exasol</strong></td>
<td>EXASolution</td>
<td>A leader in 100 GB and 1000 GB TCP-H benchmark tests</td>
</tr>
<tr>
<td><strong>InfoBright</strong></td>
<td>Brighthouse</td>
<td>Data pack storage approach for data-optimized compression for smaller footprint. Queries focus on relevant packs for faster performance.</td>
</tr>
<tr>
<td><strong>ParAccel</strong></td>
<td>ParAccel Analytic Database</td>
<td>Combines column-store architecture with in-memory analysis capabilities for fast performance. Offers &quot;drop-in accelerator&quot; configurations on top of popular relational databases.</td>
</tr>
<tr>
<td><strong>Sand Technology</strong></td>
<td>Sand/DNA Software</td>
<td>Extensions available for Oracle, IBM DB2, SAP NetWeaver BI</td>
</tr>
<tr>
<td><strong>Sybase</strong></td>
<td>Sybase IQ</td>
<td>Category leader and early pioneer with 1,000-plus customers</td>
</tr>
<tr>
<td><strong>Vertica</strong></td>
<td>Vertica Database</td>
<td>Supports continuous data loading. Also offers appliance based on HP hardware/Red Hat Linux OS-based database appliance.</td>
</tr>
</tbody>
</table>
In-memory OLAP – BI Tools

• QlikTek / QlikView
  – 64bit in-memory OLAP
    • No pre-aggregation, no hierarchies, no cube building
    – Claims 10x compression of data in memory
    – Windows OS based and exports data to Excel

• QlikTek controls the user experience
  – Fast, easy to use UI builder (dashboards, reports, etc)
  – High quality visualization and charting
  – Has its own in-memory OLAP for performance
  – Flexibility - Business user driven, not IT experts
In-memory OLAP - servers

• Typical servers have 32GB – 64GB RAM
  – With 64bit OS that’s and good compression that’s approximately a 100GB cube in memory.

• Applix TM1 (acquired by Cognos, then IBM)
  – Includes Applix Viewer (EIS)
  – Microsoft and .NET based development
  – 64-bit capable

• Palo OLAP Server 2.0 (Open Source)
  – Free in-memory OLAP server
  – ROLAP based, no aggregates stored in memory
  – Primary front end is Microsoft Excel
  – Jedox Enterprise Spreadsheets
Polynomial MOLAP (non-matrix)

• Mathematical indexing of data relationships in polynomial form: \( aD_1 + bD_2 + cD_3 \ldots = (m_1, m_2\ldots m_n) \)

• Polynomial Strings with Fractal Compression

• Highly scalable architecture, multi terabyte support

• Multi-path filtering, Merge, Union, Dynamic or Fixed Hierarchies

• Unique Key-back feature, link to detail source allows for HOLAP drill through
Polynomial HOLAP MPP Appliance

HOLAP with drill through to detail
Multi-dimensional Aggregate Awareness

Table ↔ MOLAP ↔ MPP

Disk/CPU/Memory/Network/SQL

Hundreds of nodes

Parallel Data Loading

BI Tool
Or
End User Query

Dataupia’s Satori Server Appliance with Dynamic Aggregation option
Oracle 11g OLAP option (HOLAP)

Incoming User Query

Aggregate Awareness

Query Redirected

Load Data

ROLAP mode

MOLAP mode

Predefined Summaries

Express

Synchronized

Synchronized

Synchronized

MV

MV

Dim

Dim

Dim

Dim

Fact

Detail


Tailored Delivery is an approach getting much attention

A NEW MODEL FOR USERS
Tailored Delivery

- A new self-service BI model introduced by TDWI’s Wayne Eckerson, May 2008 Keynote
  - Layered information delivery system & personnel
  - Central group that generates standard set of “interactive reports”
  - An interactive, information sandbox
  - Each contains about 12 metrics and 20 dimensions
  - Users think they are performing “ad hoc” but it’s predefined
  - Addresses 80% of questions for 80% of users
  - Each “interactive report” can replace dozens to hundreds of existing reports
  - Simple, intuitive interfaces to apply filters and sorts
Performance Dashboard

Functionality

Monitor

Graphical Data

Summarized Data

Analyze

Detailed Data

Drill thru

Users

Managers

Analysts

Workers
Evolution

Functionality

Monitor

Analyze

Drill thru

Adjacent Applications

Plan

Adv. Analytics

Decide & Act

Graphical Data

Summarized Data

Detailed Data

MAD-PAD Framework
Tailored BI – Next Gen Technologies

**Roles**

- **Business Users:** Self-Service data sets
  - Intuitive Rich Interface

- **Power Users (Bus/IT):** Self-Service Master
  - Multi-dimensional data sets

- **ETL Developers (IT):**
  - Integrate, Cleanse,
  - Accurately model
  - long term detailed data sets

**Next Gen Technology**

- **Graphical Data:** Visual BI tools with embedded in-memory OLAP or Excel connected
- **Summarized Data:** Very large scale Polynomial MOLAP with Drill thru HOLAP
- **Detailed Data:** Detailed EDW in 3NF and ROLAP
  - MPP based,
  - Appliance based,
  - Columnar possibly

**MAD-PAD Framework**
Master multi-dimensional data sets

Operations
Monitor & Analyze Today’s Activities (ongoing)

Analyst
Frame a question: What happened last month? (adhoc)
User defines and builds MOLAP as needed

Marketing
Analyze History by Attribute (project)
User drills to detail data

Detail:
Attribute
Attribute
Attribute
...

Current
Historical

DRILL

DW Detail/MPP
Tailored BI with Next Gen OLAP

• MPP database architecture used as DW
  – Vast amounts of historical data online
  – Lowest detail data stored online

• MPP database leveraged as a foundation for
  – Parallel unloading of dimensional data

• Users empowered to dynamically define
  – Multi-dimensional aggregates
Predictions for

THE FUTURE
Primary industry drivers

Start with what we do know:

• Moore’s law (more speed/capacity for your $$)
• Commodity hardware (blades)
• Virtualization increasing server efficiencies
• Software maturity leads to open source

• Data explosion
• More data needed to make business decisions
• Need to make decisions closer to real-time
BI search gets incorporated

Monitoring

Analyze

Drill thru & Mine
Detail Data everywhere

Corporate documents

Internet:
Industry
Finance
News

Detail DW
Multi-hybrid models

Monitoring

Intuitive, rich, easy BI tools
More Powerful clients

Analyze

Analytics Optimized for:
- Real-time (in-memory)
- Variance & Write-back (small fast disk)
- Massive Aggregations (DW Appliances)

HOLAP Drill Thru

Detail DW
MPP or DWA

MOLAP Aware
Polynomial
Cloud Computing

Internal

Monitoring

Internal Servers

Internal Desktops, Laptops

External in a Cloud

Analyze

Detail Data

$ Transfer (Load Data)

$ Compute (Execute SQL)

$ Transfer (Receive Data)
QUESTIONS & DISCUSSION
Contact Information

• If you have further questions or comments:

  John O’Brien
  Zukeran Technologies Corp

  jobrien@ZukeranTechnologies.com
  (720) 841-6683