Data-Warehouse-, Data-Mining- und OLAP-Technologien

Chapter 5: Online Analytic Processing

Bernhard Mitschang
Universität Stuttgart

Winter Term 2014/2015
Overview

• OLAP
  ▪ Introduction
  ▪ Operations
  ▪ Characteristics

• Storage of OLAP cubes
  ▪ Relational vs. Multidimensional
  ▪ Multidimensional Arrays
  ▪ Sparse Cubes
  ▪ Multidimensional Query Language

• Architecture
  ▪ MOLAP, ROLAP, HOLAP
OLAP

- **Online Analytic Processing**
- Technologies and tools that support (ad-hoc) analysis of multi-dimensionally aggregated data
- Individual analysis is supported, i.e., the user is not restricted to available standard reports/analysis
- Graphical user interface is available for analysis specification
- Knowledge of a query language or programming language is not required
- Result information is given graphically and made available for incorporation into other applications
- Users: Analysts, Manager, “knowledge worker”
- Typical analysis scenarios:
  - Multi-dimensional views, e.g. turnover per product group and month
  - Comparisons, e.g. turnover in Q4 compared to that of Q3
  - Ranking, e.g. top 10 product in a certain group ranked by turnover
Online Analytic Processing

OLAP

- Defining OLAP reports
  - select facts
  - select dimensions
  - define filters
  - define presentation

Top 10 fruit and vegetables

<table>
<thead>
<tr>
<th>Rank</th>
<th>Produkt</th>
<th>Turnover ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>potatoes</td>
<td>210000</td>
</tr>
<tr>
<td>2</td>
<td>carrots</td>
<td>205000</td>
</tr>
<tr>
<td>3</td>
<td>celery</td>
<td>190000</td>
</tr>
<tr>
<td>3</td>
<td>tomatoes</td>
<td>190000</td>
</tr>
<tr>
<td>5</td>
<td>kiwi fruit</td>
<td>150000</td>
</tr>
<tr>
<td>6</td>
<td>strawberry</td>
<td>145000</td>
</tr>
<tr>
<td>7</td>
<td>spinach</td>
<td>142000</td>
</tr>
<tr>
<td>8</td>
<td>zucchini</td>
<td>95500</td>
</tr>
<tr>
<td>9</td>
<td>lettuce</td>
<td>94000</td>
</tr>
<tr>
<td>10</td>
<td>blackberry</td>
<td>92000</td>
</tr>
</tbody>
</table>

Turnover per product and month

Turnover Q3 vs. Q4
Multidimensional Model

Fact Data (sales)

„Cube“ Metaphor

Product

Location

Time

Country

2002
• **Slice:**
  - restrict one dimension to a range of values

• **Dice:**
  - restrict several dimensions to a range of values
  - results in a sub-cube

• **Example:** Analysis of a certain product family.
Roll-up and Drill-down

- Roll-up (drill-up):
  - summarize data by climbing up hierarchy or by dimension reduction
- Drill-down (roll-down):
  - reverse of roll-up
  - from higher level summary to lower level summary or detailed data, or introducing new dimensions
Pivot and Rotate

- **Pivot:**
  - reorient the cube
  - visualization
  - 3D to series of 2D planes
OLAP Operations: Overview

- Typical OLAP operations (explained in a general manner):
  - **Roll up** (drill-up): summarize data
    - by climbing up hierarchy or by dimension reduction
  - **Drill down** (roll down): reverse of roll-up
    - from higher level summary to lower level summary or detailed data, or introducing new dimensions
  - **Slice and dice**: project and select
  - **Pivot** (rotate):
    - reorient the cube, visualization, 3D to series of 2D planes.
  - **Other operations**
    - **drill across**: involving (across) more than one fact table
    - **drill through**: through the bottom level of the cube to its back-end relational tables (using SQL)
OLAP Product Evaluation Rules

**Basic Features**
- R1: multi-dimensional conceptual view
- R10: intuitive data manipulation
- R3: accessibility
- N: batch extraction vs. interpretive
- N: OLAP analysis models
- R5: client-server architecture
- R2: transparency
- R8: multi-user support

**Reporting Features**
- R11: flexible reporting
- R4: consistent reporting performance
- R7: dynamic sparse matrix handling

**Dimension Control**
- R6: generic dimensionality
- R12: unlimited dimensions and aggregation levels
- R9: unrestricted cross-dimensional operations

**Special Features**
- N: treatment of non-normalized data
- N: storing OLAP results: keeping them separate from source data
- N: extraction of missing values
- N: treatment of missing values

R1 - R12: original rules
N: additional rules

Source: [CCS93]
# FASMI Test

| FAST | • deliver most responses within about five seconds  
|      | • simplest analysis taking no more than one second  
|      | • very few taking more than 20 seconds |
| ANALYSIS | • cope with any business logic and statistical analysis that is relevant for applications and users  
|          | • allow users to define new ad-hoc calculations without programming |
| SHARED | • confidentiality  
|         | • concurrent update locking if multiple write access is needed |
| MULTIDIMENSIONAL | • multidimensional conceptual view of data  
|                | • support for hierarchies and multiple hierarchies |
| INFORMATION | • handle huge amounts of input data |

Overview

• OLAP
  ▪ Introduction
  ▪ Operations
  ▪ Characteristics

• Storage of OLAP cubes
  ▪ Relational vs. Multidimensional
  ▪ Multidimensional Arrays
  ▪ Sparse Cubes
  ▪ Multidimensional Query Language

• Architecture
  ▪ MOLAP, ROLAP, HOLAP
Relational Storage of OLAP Cubes

• Mapping the cube view to a star- or snowflake-schema
• Information requests of the users have to be mapped to the relational schema (see 'sequence of typical star queries')
• Result tables have to be mapped to the cube structure before they are presented to the user
**Sequence of typical star queries (1)**

**Information request:**
Which are the top products whose number of sold pieces in the months chosen by the user compared to the respective month ago has increased most?

```sql
INSERT INTO A1 (orderyearkey, ordermonthkey, partkey, sumquantity)
SELECT od.orderyearkey, od.ordermonthkey, lo.partkey, SUM(lo.quantity)
FROM lineitem_orders lo, orderday od
WHERE od.orderdate = lo.orderdate
AND od.ordermonthkey IN (199401, 199402)
GROUP BY od.orderyearkey, od.ordermonthkey, lo.partkey;
```

Number of sold parts in January and February 1994
**Information request:**
Which are the top products whose number of sold pieces in the months chosen by the user compared to the respective month ago has increased most?
Sequence of typical star queries (3)

Information request:
Which are the top products whose number of sold pieces in the months chosen by the user compared to the respective month ago has increased most?
Information request:
Which are the top products whose number of sold pieces in the months chosen by the user compared to the respective month ago has increased most?
Multidimensional Storage of OLAP Cubes

- Allows to directly store the cells of a data cube in a n-dimensional array
- Avoids mapping between cube view and relational schema
- May result in sparse cubes
- Multidimensional query language needed
### Multidimensional Database Systems

- Allow to directly store the cells of a data cube in a n-dimensional array.

<table>
<thead>
<tr>
<th></th>
<th>single cube</th>
<th>many cubes</th>
</tr>
</thead>
<tbody>
<tr>
<td>single measure per cube</td>
<td></td>
<td>• relevant dimensionality for each measure</td>
</tr>
<tr>
<td>multiple measures per cube</td>
<td>• sparse dimensions likely</td>
<td>• direct mapping of the conceptual model</td>
</tr>
</tbody>
</table>

- Many proprietary implementations of storage structure:
  - similar to common index structures
Multidimensional Arrays

- Dimensions $D_1, \ldots, D_n$
- Data cube with $|D_1| \times |D_2| \times \ldots \times |D_n|$ cells
- Index of cell $(x_1, x_2, \ldots, x_n)$
  \[
  = x_1 + (x_2 - 1) \cdot |D_2| + (x_3 - 1) \cdot |D_1| \cdot |D_2| + \ldots + (x_n - 1) \cdot |D_1| \cdot \ldots \cdot |D_{n-1}|
  \]
  \[
  = 1 + \sum_{i=1}^{n} (x_i - 1) \cdot \prod_{j=1}^{i-1} |D_i|
  \]

- Example:
  - Dimension 1: Product
  - Dimension 2: Month
  - Which cell stores data for product C in April 2005?
Query Processing in Multidimensional Arrays

- Query processing:
  - determine index of cells
  - read pages/blocks for these cells into main memory

- Query performance depends on the number of pages to be read.

- Example:
  - How many blocks need to be read to get all cells on product A?
  - How many blocks need to be read to get all cells for February 2005?

- Order of dimensions is significant for query performance.

<table>
<thead>
<tr>
<th></th>
<th>A (1)</th>
<th>B (2)</th>
<th>C (3)</th>
<th>D (4)</th>
<th>E (5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan 05</td>
<td>1</td>
<td>6</td>
<td>11</td>
<td>16</td>
<td>5</td>
</tr>
<tr>
<td>Feb 05</td>
<td>2</td>
<td>7</td>
<td>12</td>
<td>17</td>
<td>10</td>
</tr>
<tr>
<td>Mar 05</td>
<td>3</td>
<td>8</td>
<td>13</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>Apr 05</td>
<td>4</td>
<td>9</td>
<td>14</td>
<td>19</td>
<td>20</td>
</tr>
</tbody>
</table>
Multidimensional Partitioning

- Dimensions $D_1, \ldots, D_n$
- Dimension values $1 \ldots d_i$ for each dimension $D_i$.
- Partition $b_1, \ldots, b_m$ as
  
  \[ b_1 = [l_{1,1}:u_{1,1}, \ldots, l_{1,n}:u_{1,n}] \]
  
  \[ \ldots \]
  
  \[ b_m = [l_{m,1}:u_{m,1}, \ldots, l_{m,n}:u_{m,n}] \]

- regular partitioning: same value range in dimension $D_i$ for each partition $b_j$.

- irregular partitioning: partition-specific value ranges
Multidimensional Partitioning

- Automatic partitioning:
  - system automatically defines the partitioning
  - goals:
    - identify sparse dimensions
    - efficient query processing
- Partitioning based on dimension semantics:
  - e.g. partitioning according to time series
- user-defined partitioning:
  - explicit specification based on
    - value ranges
    - dimensions
- Storage of partitions:
  - relational: coordinates of cells are stored as primary key in a table
  - array: cells are stored in an array (as shown before)
Sparse Cubes

- A cube may contain empty cells.
- Density of a cube:
  \[
  \text{Density} = \frac{\text{number of defined cells}}{\text{number of all cells}}
  \]
- N-dimensional array is efficient for dense cubes.
- Sparse cubes need further optimizations:
  - don't store empty pages/blocks
  - multidimensional partitioning + two storage levels
- Two storage levels:
  - first level:
    - index structure for sparse dimensions
    - index structures like B-trees, Grid, Hashing
  - second level:
    - n-dim. array for dense dimensions
    - compressed arrays
Multidimensional Query Language

- Query language that includes specific features for multidimensional data:
  - access to cubes
  - access to dimensions
  - aggregation of measure
  - restrictions on dimensions
  - selection of subcubes
  - set of functions for the manipulation of data
- No standard available
- Most tools provide queries based on the information users requested by means of a graphical user interface

- Example: MDX (MultiDimensional Expression)
  - published in 1998
  - part of Microsofts OLE DB for OLAP
  - OLE DB provides COM interfaces for access to various data sources
  - supports the definition and manipulation of multidimensional objects and data (DML and DDL statements)
MDX

- **Basic syntax:**
  
  ```
  SELECT [<axis_specification>] [, <axis_specification>...] ]
  FROM [<cube_specification>]
  WHERE [<slicer_specification>]]
  ```

- **SELECT clause:**
  - determines the axis dimensions of an MDX SELECT statement

- **FROM clause:**
  - determines which multidimensional data source is to be used when extracting data to populate the result set

- **WHERE clause:**
  - determines which dimension or member to use as a slicer dimension
  - slicer dimension = dimension that is not assigned to an axis
  - restricts the extracting of data to a specific dimension or member
### MDX: Examples

**SELECT**
```
{ [Measures].[Unit Sales], [Measures].[Store Sales] } ON COLUMNS,
{ [Time].[1997], [Time].[1998] } ON ROWS
```
**FROM**
```
Sales
```
**WHERE**
```
([Store].[USA].[CA])
```

- Specifies that:
  - two measures should be presented in columns
  - values for two years should be presented in rows
  - only stores in CA should be included

### WHERE clauses

- **WHERE tuple** uniquely identifies a section in the cube (subcube)
- if multiple tuples are specified (set) result cells in every tuple along the set will be aggregated

```
WHERE ( [Route].[All], [Time].[1st half] )
WHERE { ([Time].[1st half], [Route].[nonground]), ([Time].[1st half], [Route].[ground])
```
Overview

- OLAP
  - Introduction
  - Operations
  - Characteristics
- Storage of OLAP cubes
  - Relational vs. Multidimensional
  - Multidimensional Arrays
  - Sparse Cubes
  - Multidimensional Query Language
- Architecture
  - MOLAP, ROLAP, HOLAP
Architecture

- Different options based on:
  - storage of OLAP data
    - relational database
    - multidimensional database
    - files on the client
  - processing of OLAP data
    - processing SQL on the server
    - processing multidimensional queries on the server
    - processing multidimensional queries on the client
MOLAP, ROLAP, HOLAP

• MOLAP
  - data resides in a multidimensional DBMS
  - multidimensional engine (OLAP server) provides access

• ROLAP
  - data resides in a relational DBMS
  - OLAP server provides SQL queries

• HOLAP
  - detailed data resides in a relational DBMS
  - aggregated data resides in a multidimensional DBMS
# Architecture: Comparison

<table>
<thead>
<tr>
<th>Pros</th>
<th>MOLAP</th>
<th>ROLAP</th>
<th>HOLAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>• short response time</td>
<td>• mature relational technology</td>
<td>• short response time for aggregated data</td>
<td></td>
</tr>
<tr>
<td>• efficient storage structure</td>
<td>• no limits on volumes of data</td>
<td>• efficient storage structure for aggregated data</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>• no limits on volumes of data</td>
<td></td>
</tr>
<tr>
<td>Cons</td>
<td></td>
<td>• increased response time</td>
<td>• increased response time</td>
</tr>
<tr>
<td>• limited performance for large volumes of data</td>
<td>• increased response time</td>
<td>• increased response time for detailed data</td>
<td></td>
</tr>
<tr>
<td>• large volumes of data on OLAP server (detailed and aggregated)</td>
<td></td>
<td>• administration</td>
<td></td>
</tr>
<tr>
<td>• preprocessing to provide OLAP cubes</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Product Overview

<table>
<thead>
<tr>
<th>Vendor</th>
<th>Product</th>
<th>Market Coverage in 2004</th>
</tr>
</thead>
<tbody>
<tr>
<td>Microsoft</td>
<td>SQL Server 2000 Analysis Services</td>
<td>27,4%</td>
</tr>
<tr>
<td>Hyperion Solutions</td>
<td>Hyperion Essbase, Hyperion Intelligence (former BRIO)</td>
<td>20,7%</td>
</tr>
<tr>
<td>Cognos</td>
<td>PowerPlay, Planning (former Adaytum Planning)</td>
<td>14,1%</td>
</tr>
<tr>
<td>Business Objects</td>
<td>BusinessObjects and Webintelligence</td>
<td>7,2%</td>
</tr>
<tr>
<td>MicroStrategy</td>
<td>MicroStrategy7i</td>
<td>7,1%</td>
</tr>
<tr>
<td>SAP</td>
<td>SAP Business Information Warehouse</td>
<td>6,0%</td>
</tr>
<tr>
<td>Oracle</td>
<td>Oracle Express, Oracle10g OLAP Option</td>
<td>3,7%</td>
</tr>
<tr>
<td>Applix</td>
<td>Applix TM1</td>
<td>3,1%</td>
</tr>
<tr>
<td>Cartesis</td>
<td>Cartesis Magnitude</td>
<td>3,1%</td>
</tr>
</tbody>
</table>

(www.olapreport.com)
Market Overview

Worldwide total OLAP market

Source: The OLAP Report

www.olapreport.com

OLAP market share trend

Source: The OLAP Report

www.olapreport.com

(www.olapreport.com)
Summary

• OLAP: Technologies and tools that support (ad-hoc) analysis of multi-dimensionally aggregated data
• Basic Operations:
  ▪ Slice and Dice, Roll-up and Drill-down, Pivot
• Main characteristics of OLAP:
  ▪ Fast, Analysis, Shared, Multidimensional, Information
• Storage options:
  ▪ relational database system
  ▪ multidimensional db (n-dimensional arrays, m-dim. query language)
• Architectural options:
  ▪ ROLAP, MOLAP, HOLAP
Papers