MTAT.03.183 Data Mining

OLAP

Siim Karus
siim.karus@ut.ee

Spring 2017

https://courses.cs.ut.ee/2017/dm/
Siim Karus

- PhD in Computer Science
- Microsoft Student Partner 2007-2009 (MSP)
- Author of Estonian freeware portal VabaVaraVeeb
  - http://vabavara.eu
- More than 10 years of industry experience as programmer, information systems’ auditor, architect and business intelligence developer.

05.05.17

MTAT.03.183 Data Mining
Seminar on Business Intelligence

• Introduce the practices and methods used for extracting and presenting relevant information for decision-making out of databases.

• Give background on data integration (including warehousing), data analysis in databases and data presentation.
Business Intelligence

“The ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal.”

– Hans Peter Luhn (1958, IBM)
“A broad category of applications and technologies for gathering, storing, analyzing, sharing and providing access to data to help enterprise users make better business decisions.”
– Gartner
Visual Mining (cubes, dimensions, partitions)
Data Mining (choice of algorithms)
Visual Machine Learning (learning from user interactions with results)
Social BI
Interpretation

DATA ANALYSIS
On-line analytical processing

- Optimised for information queries (SELECT)
- Based on data warehouse principles (multiple data sources)
- Fact-dimension logical view
  - ROLAP – relational-OLAP – multi-dimensional queries are translated into relational queries from relational database
  - MOLAP – multidimensional-OLAP – the data is stored in multi-dimensional storage
  - HOLAP – hybrid-OLAP
Data Cubes
Dimension design

- **Regular (Simple) Dimensions (Star schema)**
- **Referenced Dimensions (Snowflake schema)**
- **Fact (Many-to-Many) Dimension**
Example
## Dimension hierarchies

### Hierarchies

<table>
<thead>
<tr>
<th>Semesters</th>
<th>Calendar Weeks</th>
<th>Calendar Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Academic Year</td>
<td>- Year</td>
<td>- Year</td>
</tr>
<tr>
<td>- Semester</td>
<td>- Week</td>
<td>- Half Of Year</td>
</tr>
<tr>
<td>- Month</td>
<td>- Weekday</td>
<td>- Month</td>
</tr>
<tr>
<td>- Day</td>
<td>- Day</td>
<td>- Day</td>
</tr>
<tr>
<td>- Date</td>
<td>- Date</td>
<td>- Date</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Academic Weeks</th>
</tr>
</thead>
<tbody>
<tr>
<td>- Academic Year</td>
</tr>
<tr>
<td>- Semester</td>
</tr>
<tr>
<td>- Academic Week</td>
</tr>
<tr>
<td>- Weekday</td>
</tr>
<tr>
<td>- Day</td>
</tr>
<tr>
<td>- Date</td>
</tr>
</tbody>
</table>

To create a new hierarchy, drag an attribute here.
Dimension Granuality
Dimension changes

• Recomputed dimension -> recompute cube

• Dimension changes
  – Speed of change (slow, fast)
  – Type of change (appending, insertion, modification)
  – Reforms (e.g. Organisational structure changes)
Measures

- **Simple aggregations**
  - Min
  - Max
  - Average
  - Sum
  - Count

- **Complex aggregations**
  - Difference with previous period
  - Conditional sum or count

- **Calculation types**
  - Precomputed vs computed during runtime
  - Over visible nodes vs over all nodes
Partitions

• Split cube by dimension values

• Partitioning:
  – Different data sources
  – Different storage policies (e.g. Operative non-cached data ROLAP partition and historic cached data MOLAP partition)
  – Read-only vs. Read-write partitions
Perspectives

• Subsets of cubes (not necessarily subcubes)

• Purpose:
  – show cube data relevant to different stakeholders
KPIs

- Key Performance Indicators
  - Value
  - Target
  - Trend (indicator)
  - Status (indicator)
Operations

• Dimension-based
  – Roll-up (Drill-up)
  – Drill-down

• Interaction-based
  – Drill-through

• Dimensionality
  – Slice (2D) and Dice (3D)
  – Pivot (rotate)
MDX

- MDX – Multi-Dimensional eXpressions (to SQL)
- DMX – Data-Mining eXpressions (to SQL)
- DAX – Data Access eXpressions (from Excel, Power BI, self-service BI platforms)
MDX

• WITH

  <member definitions>

SELECT

  <fact definition>,

  <dimension definitions>

FROM

  <cube selector>

WHERE

  <filter expression>
MDX example

SELECT
  { [Measures].[Sales Amount], [Measures].[Tax Amount] } ON COLUMNS,
  { [Date].[Fiscal].[Fiscal Year].&[2002], [Date].[Fiscal].[Fiscal Year].&[2003] } ON ROWS
FROM
  [Adventure Works]
WHERE
  ( [Sales Territory].[Southwest] )
WITH
  MEMBER [Measures].[ParameterCaption] AS [Start Form Ref].[Study Form Id].CURRENTMEMBER.MEMBER_CAPTION
  MEMBER [Measures].[ParameterValue] AS [Start Form Ref].[Study Form Id].CURRENTMEMBER.UNIQUENAME
  MEMBER [Measures].[ParameterLevel] AS [Start Form Ref].[Study Form Id].CURRENTMEMBER.LEVEL.ORDINAL
SELECT
  {[Measures].[ParameterCaption], [Measures].[ParameterValue], [Measures].[ParameterLevel]} ON COLUMNS,
  [Start Form Ref].[Study Form Id].ALLMEMBERS ON ROWS
FROM
  [Student Semester]
OLAP Software

- Microsoft SQL Server Analysis Services
- Oracle Essbase (MOLAP only)
- Oracle OLAP – (ROLAP only)
- IBM Cognos TM1 (MOLAP only)
- IBM Cognos BI (no local or in-memory option)
- SAS OLAP Server (no local or in-memory option)
Distributed OLAP

• Druid
• Apache Kylin (eBay)
• Cubes
• Pinot (LinkedId)
OLAP clients

- ODBC
- XMLA (XML for Analysis)
- OLAP design tools
- Excel, PowerPivot
XMLA example

```xml
<soap:Envelope>
  <soap:Body>
    <Execute xmlns="urn:schemas-microsoft-com:xml-analysis">
      <Command>
        <Statement>SELECT Measures.MEMBERS ON COLUMNS FROM Sales</Statement>
      </Command>
      <Properties>
        <PropertyList>
          <DataSourceInfo/>
          <Catalog>FoodMart</Catalog>
          <Format>Multidimensional</Format>
          <AxisFormat>TupleFormat</AxisFormat>
        </PropertyList>
      </Properties>
    </Execute>
  </soap:Body>
</soap:Envelope>
```
Alternatives

- SQL PIVOT, UNPIVOT, CUBE, ROLLUP expressions
- OLTP statistics
- Read-only OLTP mirror
- SQL Cube Views (DB2)
OLAP usage in DM

- Data Mining models to populate/augment cubes with data
- Certain machine learning algorithms operate on cubes (e.g. Shopping cart analysis)