Data Warehousing
Lecture 2

Toon Calders
toon.calders@ulb.ac.be
Previous Lecture
OLAP

• OLAP = OnLine Analytical Processing
  – Online = no waiting for answers

• OLAP system = system that supports analytical queries that are dimensional in nature.

• Most data warehousing systems support OLAP functionalities
Outline

Online Analytical Processing

• Conceptual model: Data cubes

• Query languages for supporting OLAP
  – Typical data cube operations
  – SQL extensions
  – MDX

• Database Explosion Problem
Supermarket Example

- Evaluate the sales of products
  - Product cost in $
  - Customer: ID, city, state, country,
  - Store: chain, size, location,
  - Product: brand, type, ...
  - ...

Dim.  measure

Cost in $

store  customer

product
Supermarket Example

- Multi-dimensional view on data
Cross Tabulation

- Cross-tabulations are highly useful
  - Sales of clothes June → August ‘06

<table>
<thead>
<tr>
<th></th>
<th>Blue</th>
<th>Red</th>
<th>Orange</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>June</td>
<td>51</td>
<td>25</td>
<td>158</td>
<td>234</td>
</tr>
<tr>
<td>July</td>
<td>58</td>
<td>20</td>
<td>120</td>
<td>198</td>
</tr>
<tr>
<td>August</td>
<td>65</td>
<td>22</td>
<td>51</td>
<td>138</td>
</tr>
<tr>
<td>Total</td>
<td>174</td>
<td>67</td>
<td>329</td>
<td>570</td>
</tr>
</tbody>
</table>

Date: month, June → August 2006
Data Cubes

• Extension of Cross-Tables to multiple dimensions
  – Conceptual notion

<table>
<thead>
<tr>
<th></th>
<th>Blue</th>
<th>Red</th>
<th>Orange</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>June</td>
<td>51</td>
<td>25</td>
<td>158</td>
<td>234</td>
</tr>
<tr>
<td>July</td>
<td>58</td>
<td>20</td>
<td>120</td>
<td>198</td>
</tr>
<tr>
<td>August</td>
<td>65</td>
<td>22</td>
<td>51</td>
<td>138</td>
</tr>
<tr>
<td>Total</td>
<td>174</td>
<td>67</td>
<td>329</td>
<td>570</td>
</tr>
</tbody>
</table>

Dimensions

- Aggregated w.r.t. X-dim
- Aggregated w.r.t. Y-dim
- Aggregated w.r.t. X and Y

Data Points/1st level of aggregation
Data Cubes

<table>
<thead>
<tr>
<th>Product</th>
<th>1Qtr</th>
<th>2Qtr</th>
<th>3Qtr</th>
<th>4Qtr</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Country

- Ireland
- France
- Germany
Data Cubes

Aggregated over:
- country
- date
- product
Data Cubes

Aggregated over
- product & date
- date & country
- product & country
Data Cubes

<table>
<thead>
<tr>
<th>Product</th>
<th>Product Sum</th>
<th>Country</th>
<th>Country Sum</th>
</tr>
</thead>
<tbody>
<tr>
<td>TV</td>
<td></td>
<td>Ireland</td>
<td></td>
</tr>
<tr>
<td>PC</td>
<td></td>
<td>France</td>
<td></td>
</tr>
<tr>
<td>VCR</td>
<td></td>
<td>Germany</td>
<td></td>
</tr>
<tr>
<td>Product Sum</td>
<td>1Qtr</td>
<td>2Qtr</td>
<td>3Qtr</td>
</tr>
<tr>
<td>TV</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PC</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VCR</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Aggregated over all
SSAS – Data Cubes

Cube name

Measures

“key performance indicators”

Dimensions
SSAS - Dimension

- Attributes
  - Customer
    - Demographic
      - Commute Distance
      - Education
      - Gender
      - Home Owner
      - Marital Status
      - Number of Cars Owned
      - Number of Children At Home
      - Occupation
      - Total Children
      - Yearly Income
  - Location
    - City
    - Country
    - Postal Code
    - State-Provence
- Hierarchy
- Levels
• One dimension can have multiple hierarchies
• Hierarchies consist of *levels*
• Levels are in a linear order
<table>
<thead>
<tr>
<th>ALL</th>
<th>Country</th>
<th>State-Province</th>
<th>City</th>
<th>Postal code</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Australia</td>
<td>New South Wales</td>
<td>Alexandria</td>
<td>2015</td>
<td>Adriana Smith</td>
</tr>
<tr>
<td></td>
<td>Canada</td>
<td>Queensland</td>
<td>Coffs Harbour</td>
<td>2450</td>
<td>Aimee Guo</td>
</tr>
<tr>
<td></td>
<td>France</td>
<td>South Australia</td>
<td>Darlinghurst</td>
<td>2010</td>
<td>Allison R. Young</td>
</tr>
<tr>
<td></td>
<td>Germany</td>
<td>Tasmania</td>
<td>Goulburn</td>
<td>2580</td>
<td>Ann A. Sara</td>
</tr>
<tr>
<td></td>
<td>United Kingdom</td>
<td>Victoria</td>
<td>Lane Cove</td>
<td>1597</td>
<td>Antonio G. Patterson</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td>Alberta</td>
<td>Lavender Bay</td>
<td>2060</td>
<td>Ariana Stewart</td>
</tr>
<tr>
<td></td>
<td></td>
<td>British Columbia</td>
<td>Malabar</td>
<td>2036</td>
<td>Arthur Kapoor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Brunswick</td>
<td>Matraville</td>
<td>2036</td>
<td>Barbara W. Lal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Manitoba</td>
<td>Milsons Point</td>
<td>2061</td>
<td>Bobby D. Saunders</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Ontario</td>
<td>Newcastel</td>
<td>2300</td>
<td>Brianna J. Johnson</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quebec</td>
<td>North Ryde</td>
<td>2113</td>
<td>Bruce G. Madan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Charente-Maritime</td>
<td>North Sydney</td>
<td>2055</td>
<td>Bryant L. Perez</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Essonne</td>
<td>Port Macquarie</td>
<td>2444</td>
<td>Carla D. Madan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Garonne (Haute)</td>
<td>Rhodes</td>
<td>2138</td>
<td>Carlos Edwards</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gers</td>
<td>Silverwater</td>
<td>2264</td>
<td>Carly Anand</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hauts de Seine</td>
<td>Springwood</td>
<td>2777</td>
<td>Cedric Liu</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Loir et Cher</td>
<td>St. Leonards</td>
<td>2065</td>
<td>Clarence Xu</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Sydney</td>
<td>1002</td>
<td>Colin Chavez</td>
</tr>
</tbody>
</table>
# Children, Parent

## Hierarchy Geography on Dimension Customer

<table>
<thead>
<tr>
<th>ALL</th>
<th>Country</th>
<th>State-Province</th>
<th>City</th>
<th>Postal code</th>
<th>Customer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Australia</td>
<td></td>
<td></td>
<td></td>
<td>Adriana Smith</td>
</tr>
<tr>
<td></td>
<td>Canada</td>
<td></td>
<td></td>
<td></td>
<td>Aimee Guo</td>
</tr>
<tr>
<td></td>
<td>France</td>
<td></td>
<td></td>
<td></td>
<td>Allison R. Young</td>
</tr>
<tr>
<td></td>
<td>Germany</td>
<td></td>
<td></td>
<td></td>
<td>Ann A. Sara</td>
</tr>
<tr>
<td></td>
<td>United Kingd</td>
<td></td>
<td></td>
<td></td>
<td>Antonio G. Patterson</td>
</tr>
<tr>
<td></td>
<td>United States</td>
<td></td>
<td></td>
<td></td>
<td>Ariana Stewart</td>
</tr>
<tr>
<td></td>
<td></td>
<td>New South Wales</td>
<td></td>
<td></td>
<td>Arthur Kapoor</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Queensland</td>
<td></td>
<td></td>
<td>Barbara W. Lal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>South Australia</td>
<td></td>
<td></td>
<td>Bobby D. Saunders</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tasmania</td>
<td></td>
<td></td>
<td>Brianna J. Johnson</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Victoria</td>
<td></td>
<td></td>
<td>Bruce G. Madan</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alberta</td>
<td></td>
<td></td>
<td>Bryant L. Perez</td>
</tr>
<tr>
<td></td>
<td></td>
<td>British Columbia</td>
<td></td>
<td></td>
<td>Carla D. Madan</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Carlos Edwards</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Carly Anand</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Cedric Liu</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Clarence</td>
</tr>
</tbody>
</table>

---

**Children**

- Australia
- Canada
- France
- Germany
- United Kingdom
- United States

**Parent**

- Essonne
- Garonne (Haute)
- Gers
- Hauts de Seine
- Loir et Cher

---

**Members**

- Alexandria
- Coffs Harbour
- Darlinghurst
- Goulburn
- Lane Cove
- Lavender Bay
- Malabar
- Matraville
- Milsons Point
- New South Wales
- Queensland
- South Australia
- Tasmania
- Victoria
- Alberta
- British Columbia
- New Brunswick
Outline

Online Analytical Processing

• Conceptual model: Data cubes
• Query languages for supporting OLAP
  – Typical data cube operations
  – SQL extensions
  – MDX
• Database Explosion Problem
## Pivoting

- Change the dimensions that are “displayed”; select a cross-tab.
  - look at the cross-table for product-date
  - display cross-table for date-customer

<table>
<thead>
<tr>
<th>Sales</th>
<th>Date</th>
<th>1st sem</th>
<th>2nd sem</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ireland</td>
<td></td>
<td>20</td>
<td>23</td>
<td>43</td>
</tr>
<tr>
<td>France</td>
<td></td>
<td>126</td>
<td>138</td>
<td>264</td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td>56</td>
<td>48</td>
<td>104</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>202</td>
<td>209</td>
<td>411</td>
</tr>
</tbody>
</table>
Browsing a Cube
Browsing a Cube
Slicing

• Select a part of the cube by restricting one or more dimensions to some values
Browsing a Cube
Drill-down and Roll-Up

• Change level to a descendant in the hierarchy
  – city $\rightarrow$ store
  – country $\rightarrow$ cities
  – product type $\rightarrow$ product

• Roll-up = inverse operation

• Drill-through:
  – go back to the original, individual data records
## Browsing a Cube

In the screenshot, we see a Microsoft SQL Server Analysis Services (SSAS) cube named "Adventure Works" being browsed. The cube is opened in the MDXQuery6.mdl...p-calendars.mdl interface. The cube is being browsed through a filter on the Calendar Quarter dimension.

The filter is set on the Calendar Quarter dimension to year 2007. The filter is applied to the City dimension, specifically to the cities of Saint Ouen, Les Ulis, and Morangis.

### Data Table

<table>
<thead>
<tr>
<th>Calendar Quarter</th>
<th>City</th>
<th>Customer Count</th>
<th>Internet Order Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 CY 2007</td>
<td>Saint Ouen</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Les Ulis</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Morangis</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Q2 CY 2007</td>
<td>Saint Ouen</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Les Ulis</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Morangis</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Q3 CY 2007</td>
<td>Saint Ouen</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Les Ulis</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Morangis</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>Q4 CY 2007</td>
<td>Saint Ouen</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Les Ulis</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Morangis</td>
<td>7</td>
<td>10</td>
</tr>
</tbody>
</table>

The data shows the customer count and internet order count for each city across different calendar quarters for the year 2007.
### Browsing a Cube

#### Measure Group:
- <All>
  - Adventure Works
    - Measures
      - Internet Customers
        - Customer Count
      - Internet Orders
        - Internet Order Count
      - Internet Sales
      - Reseller Sales
      - Sales Orders
      - Sales Summary
      - Sales Targets
    - Customer
      - Demographic
      - Location
      - Sets
      - Customer Geography
        - Member Properties
          - Australia
          - Canada
          - France
          - Germany
          - United Kingdom
          - United States
        - State-Province
        - City
        - Postal Code
        - Customer
      - Country
        - Canada
        - France
        - Germany
        - United Kingdom
        - United States
      - December 2007
        - December 1, 2007
        - December 2, 2007
        - December 3, 2007
        - December 4, 2007
        - December 5, 2007
        - December 6, 2007
        - December 7, 2007
        - December 8, 2007
        - December 9, 2007
        - December 10, 2007
        - December 11, 2007
        - December 12, 2007
        - December 13, 2007
        - December 14, 2007
        - December 15, 2007
        - December 16, 2007
        - December 17, 2007
        - December 18, 2007
        - December 19, 2007
        - December 20, 2007
        - December 21, 2007
        - December 22, 2007
        - December 23, 2007
        - December 24, 2007
        - December 25, 2007
        - December 26, 2007
        - December 27, 2007
        - December 28, 2007
        - December 29, 2007
        - December 30, 2007
        - December 31, 2007
      - Q1 CY 2008
      - Q2 CY 2008
      - Q3 CY 2008
      - Q4 CY 2008
      - Total

#### Perspective:
- Adventure Works
- Language: Default

#### Dimension:
- Customer
  - Customer Geography
    - City

#### Table:
<table>
<thead>
<tr>
<th>Calendar Quarter</th>
<th>Date</th>
<th>Customer Count</th>
<th>Internet Order Count</th>
<th>Customer Count</th>
<th>Internet Order Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q1 CY 2005</td>
<td>December 2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Q4 CY 2005</td>
<td>December 2</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Q1 CY 2006</td>
<td>December 2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Q2 CY 2006</td>
<td>December 2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Q3 CY 2006</td>
<td>December 2</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Q4 CY 2006</td>
<td>December 2</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Q1 CY 2007</td>
<td>December 2</td>
<td>2</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>Q2 CY 2007</td>
<td>December 2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Q3 CY 2007</td>
<td>December 2</td>
<td>3</td>
<td>17</td>
<td>18</td>
<td>18</td>
</tr>
<tr>
<td>Q4 CY 2007</td>
<td>December 2</td>
<td>3</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>November 2007</td>
<td>December 2</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>Q1 CY 2008</td>
<td>December 2</td>
<td>7</td>
<td>7</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>Q2 CY 2008</td>
<td>December 2</td>
<td>4</td>
<td>4</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Q3 CY 2008</td>
<td>December 2</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Q4 CY 2008</td>
<td>December 2</td>
<td>21</td>
<td>31</td>
<td>90</td>
<td>140</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>7</td>
<td>7</td>
<td>26</td>
<td>31</td>
</tr>
<tr>
<td>Q1 CY 2008</td>
<td></td>
<td>7</td>
<td>7</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>Q2 CY 2008</td>
<td></td>
<td>4</td>
<td>4</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Q3 CY 2008</td>
<td></td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td>21</td>
<td>31</td>
<td>90</td>
<td>140</td>
</tr>
</tbody>
</table>
Roll-Up
Roll-Up

Product: TV, PC, VCR

Date: 1st semester, 2nd semester, sum

Country: Ireland, France, Germany, sum

sum
### Browsing a Cube

#### Adventure Works [Browse]

**Measure Group:**
- Adventure Works
  - Measures
    - Internet Customers
      - Customer Count
    - Internet Orders
      - Internet Order Count
    - Internet Sales
    - Reseller Sales
    - Sales Orders
    - Sales Summary
    - Sales Targets
  - Customer
    - Demographic
    - Location
    - Sets
    - Customer Geography
      - Member Properties
        - Australia
        - Canada
        - France
        - Germany
        - United Kingdom
        - United States
      - State-Province
      - City
      - Postal Code
      - Customer

**Dimension:**
- Customer Geography
  - City

**Operator:**
- Equal

**Drop Filter Fields Here**

<table>
<thead>
<tr>
<th>Calendar Quarter</th>
<th>Month</th>
<th>Date</th>
<th>Customer Count</th>
<th>Internet Order Count</th>
<th>Customer Count</th>
<th>Internet Order Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Q3 CY 2005</td>
<td>October</td>
<td>2007</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Q4 CY 2005</td>
<td>November</td>
<td>2007</td>
<td>3</td>
<td>3</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Q1 CY 2006</td>
<td>December</td>
<td>2007</td>
<td>7</td>
<td>7</td>
<td>13</td>
<td>14</td>
</tr>
<tr>
<td>Q2 CY 2006</td>
<td>December</td>
<td>2007</td>
<td>26</td>
<td>26</td>
<td>31</td>
<td>31</td>
</tr>
<tr>
<td>Q3 CY 2006</td>
<td>December</td>
<td>2007</td>
<td>23</td>
<td>23</td>
<td>24</td>
<td>24</td>
</tr>
<tr>
<td>Q4 CY 2006</td>
<td>December</td>
<td>2007</td>
<td>4</td>
<td>4</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Q1 CY 2008</td>
<td>December</td>
<td>2007</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Q2 CY 2008</td>
<td>December</td>
<td>2007</td>
<td>4</td>
<td>4</td>
<td>36</td>
<td>36</td>
</tr>
<tr>
<td>Q3 CY 2008</td>
<td>December</td>
<td>2007</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Grand Total</td>
<td></td>
<td></td>
<td>21</td>
<td>31</td>
<td>90</td>
<td>140</td>
</tr>
</tbody>
</table>
Dicing

• Roll-up on multiple dimensions at once
Outline

Online Analytical Processing

• Conceptual model: Data cubes

• Query languages for supporting OLAP
  – Typical data cube operations
  – SQL extensions
  – MDX

• Database Explosion Problem
Extended Aggregation

• SQL-92 aggregation quite limited
  – Many useful aggregates are either very hard or impossible to specify
    • Data cube
    • Complex aggregates (median, variance)
    • binary aggregates (correlation, regression curves)
    • ranking queries (“assign each student a rank based on the total marks”)

• SQL:1999 adds several OLAP extensions
  – Group by cube/by rollup
## Representing the Cube

<table>
<thead>
<tr>
<th>Sales</th>
<th>Date</th>
<th>1st sem</th>
<th>2nd sem</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Country</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ireland</td>
<td>Ireland</td>
<td>20</td>
<td>23</td>
<td>43</td>
</tr>
<tr>
<td>France</td>
<td>France</td>
<td>126</td>
<td>138</td>
<td>264</td>
</tr>
<tr>
<td>Germany</td>
<td>Germany</td>
<td>56</td>
<td>48</td>
<td>104</td>
</tr>
<tr>
<td>Total</td>
<td>Total</td>
<td>202</td>
<td>209</td>
<td>411</td>
</tr>
</tbody>
</table>
Representing the Cube

- Special value « null » is used:

<table>
<thead>
<tr>
<th>Date</th>
<th>Country</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st semester</td>
<td>Ireland</td>
<td>20</td>
</tr>
<tr>
<td>1st semester</td>
<td>France</td>
<td>126</td>
</tr>
<tr>
<td>1st semester</td>
<td>Germany</td>
<td>56</td>
</tr>
<tr>
<td>1st semester</td>
<td>null</td>
<td>202</td>
</tr>
<tr>
<td>2nd semester</td>
<td>Ireland</td>
<td>23</td>
</tr>
<tr>
<td>2nd semester</td>
<td>France</td>
<td>138</td>
</tr>
<tr>
<td>2nd semester</td>
<td>Germany</td>
<td>48</td>
</tr>
<tr>
<td>2nd semester</td>
<td>null</td>
<td>209</td>
</tr>
<tr>
<td>null</td>
<td>Ireland</td>
<td>43</td>
</tr>
<tr>
<td>null</td>
<td>France</td>
<td>264</td>
</tr>
<tr>
<td>null</td>
<td>Germany</td>
<td>104</td>
</tr>
<tr>
<td>null</td>
<td>null</td>
<td>411</td>
</tr>
</tbody>
</table>
Group by Cube

• group by cube:

```sql
select item-name, color, size, sum(number)
from sales
group by cube(item-name, color, size)
```

Computes the union of eight different groupings of the `sales` relation:

```plaintext
{ (item-name, color, size),
  (item-name, color),
  (item-name, size),
  (color, size),
  (item-name),
  (color),
  (size),
  ( ) }
```
Group by Cube

• Relational representation of the date-country-sales cube can be computed as follows:

```sql
select semester as date, country, sum(sales)
from sales
group by cube(semester, country)
```

Instead of:

```sql
select semester as date, country, sum(sales)
from sales
group by semester, country
UNION select null as date, country, sum(sales)
from sales
UNION select semester as date, null as country, sum(sales)
from sales
UNION select null as date, null as country, sum(sales)
from sales
```
Group by Rollup

• rollup construct generates union on every prefix of specified list of attributes

```sql
select country, province, city, sum(number)
from sales
group by rollup(country, province, city)
```

Generates union of groupings:

```
{(country, province, city), (country, province),
 (country), ( ) }
```

Useful when there is a hierarchy between items

e.g., `group by (province)` does not make sense in the presence of `group by (country, province)`
Group by Cube & Rollup

• Multiple rollups and cubes can be used in a single group by clause

```sql
select country, province, city,
category, product,
sum(number) from sales
group by rollup(country, province, city),
rollup(category, product)
```

Generates 12 groups; all combinations of:

{(country, province, city), (country, province),
(country), ()}

and

{(category, product), (category),()}

Outline

Online Analytical Processing

• Conceptual model: Data cubes

• Query languages for supporting OLAP
  – Typical data cube operations
  – SQL extensions
  – MDX

• Database Explosion Problem
MDX

- Multidimensional Expressions (MDX) is a query language for cubes
  - Supported by many data warehousing systems
    - MS SQL Server, SAS OLAP Server, drivers for MDX for Oracle OLAP
  - Works on cubes, generates Pivot Tables

```
SELECT { [Measures].[Store Sales] } ON COLUMNS,
       { [Date].[2002], [Date].[2003] } ON ROWS
FROM Sales
WHERE ( [Store].[USA].[CA] )

SELECT { continent.[Europe], continent.[Asia] } ON Axis(0),
       { Product.[Computers], Product.[Printers] } ON Axis(1),
       { Years.[1996], Years.[1997] } ON Axis(2)
FROM Sales
```
Outline

Online Analytical Processing
• Data Warehouses
• Conceptual model: Data cubes
• Query languages for supporting OLAP
  – SQL extensions
  – MDX
• Database Explosion Problem
Three-Tier Architecture

- **Data Sources**
  - Operational DBs
  - other sources

- **Data Storage**
  - Metadata
  - Extract Transform Load Refresh
  - Data Marts
  - OLAP Engine
  - ROLAP Server
  - Data Marts

- **OLAP Engine**
  - OLAP Server
  - Monitor & Integrator

- **Front-End Tools**
  - Query/Reporting
  - Data Mining
  - Analysis
Implementation

• To make query answering more efficient: consolidate (materialize) all aggregations

• Early implementations used a multidimensional array.
  – Fast lookup: cell(prod. p, date d, prom. pr):
    • look up index of p1, index of d, index of pr:
      index = (p x D x PR) + (d x PR) + pr

• Obvious problem: sparse data
  – easy to solve, though;
    • binary search tree, hash table, ...

• Nevertheless: very quickly people were confronted with the Data Explosion Problem
Outline

Online Analytical Processing

• Conceptual model: Data cubes
• Query languages for supporting OLAP
  – Typical data cube operations
  – SQL extensions
  – MDX
• Database Explosion Problem
Data Explosion Problem

• Why?
  – n dimensions, every dimension has d values
    • d^n possible tuples.
  – Number of cells in the cube: (d+1)^n
    • Only a factor d increase

• However, most data is not dense, but sparse
  – not all d^n tuples are there in the source data.

Example: 10 dimensions with 10 values
  10 000 000 000 possibilities
  *One tuple increases the count of 2^{10} cells*
  How many for N tuples?
## Dense Cube

<table>
<thead>
<tr>
<th>Country</th>
<th>Brand</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>A</td>
<td>123</td>
</tr>
<tr>
<td>FR</td>
<td>B</td>
<td>456</td>
</tr>
<tr>
<td>BE</td>
<td>A</td>
<td>678</td>
</tr>
<tr>
<td>BE</td>
<td>B</td>
<td>254</td>
</tr>
</tbody>
</table>

![Dense Cube Diagram]
## Explosion Problem: Sparsity

<table>
<thead>
<tr>
<th>Country</th>
<th>Brand</th>
<th>Sales</th>
</tr>
</thead>
<tbody>
<tr>
<td>FR</td>
<td>A</td>
<td>123</td>
</tr>
<tr>
<td>NL</td>
<td>B</td>
<td>456</td>
</tr>
<tr>
<td>BE</td>
<td>C</td>
<td>678</td>
</tr>
<tr>
<td>US</td>
<td>D</td>
<td>254</td>
</tr>
<tr>
<td>US</td>
<td>E</td>
<td>134</td>
</tr>
</tbody>
</table>

![Graph showing sparsity](image-url)
Data Explosion Problem

• Suppose:
  – m dimensions
  – n data points
  – dimensions are i.i.d.
  – all values drawn uniformly from \{0, 1\}

• Under these settings we will analyze how the size of the cube grows with the number of dimensions
Data Explosion Problem

Size of cube w.r.t. number of data points (10 dimensions)
Data Explosion Problem

Size of cube w.r.t. number of dimensions (500 data points)
Data Explosion Problem

Logscale: Size of cube w.r.t. number of dimensions
(500 data points)
Summary

• Datawarehouses supporting OLAP for decision support

• Data Cubes as a conceptual model
  – Measurement, dimensions, hierarchy, aggregation

• Queries
  – Roll-up, Drill-down, Slice and dice, pivoting...
  – SQL:1999 extensions for supporting OLAP

• Straightforward implementation is problematic