Data Warehousing
Lecture 3

Toon Calders
toon.calders@ulb.ac.be
What have we seen last time?

• Evaluate the sales of products

  – Product cost in $

  Dim. Customer: ID, city, state, country,

  Store: chain, size, location,

  Product: brand, type, ...

  ...
What have we seen last time?

Country
- Ireland
- France
- Germany

Aggregated over all

Product
- TV
- PC
- VCR

sum

Date
- 1Qtr
- 2Qtr
- 3Qtr
- 4Qtr
- sum

sum

Country
Outline

- Dimensional fact model
  - Basic concepts
  - Extensions
- Roll-up lattice
- Special aggregation cases
- Properties of measures and aggregations

Chapter 5 of Golfarelli & Rizzi
Dimensional Fact Model

• Important to model before implement
  – Communication and Documentation
  – Facilitates maintenance and reuse

• Entity Relationship model is less suitable
  – Not focused on the dimensional model; no notion of dimension, hierarchy, ...

• We will use DFM as modeling language
Basic Concepts: Fact

• *Fact*: most specific unit of data that will be used in the analysis.
  – Usually corresponds to one or more transactions within a company
  – We will typically analyze sets of homogeneous facts; that is: facts with the same attributes

• What will be considered a fact = design choice
  – single sale; sales transaction;
    all sales of a product on a given day and shop
Examples: Fact

• On 01/01/2013 at 7:15, customer 0098745 bought product 12345 for the price of 10.95 EUR plus 20% VAT.

• On 01/01/2013, in our store “Brussels-av. Louise”, 145 items of product 01245 have been sold for an average price of 123.57 EUR.
Basic Concepts: Dimension

- *Dimension*: A fact property; a coordinate of the fact.
  - A dimension may have multiple dimensional attributes
  - Every fact corresponds to a unique combination of values for the dimensions.

- Design choice: how are the dimensional attributes grouped in dimensions
Examples: Dimension

• Fact “On 01/01/2013 at 7:15, customer 0098745 bought product 12345 for the price of 10.95 EUR plus 20% VAT.”

• Dimension Customer
  – Attributes: market segment, city, date of birth

• Dimension Date
  – Attributes: year, semester, quarter, month, day

• Dimension Product
  – Attributes: code, brand, type
Basic Concepts: Hierarchy

• Dimensions have *hierarchies*. Hierarchies express how the values of a dimension can be generalized
  – Hierarchy is a directed acyclic graph (DAG) whose nodes are dimensional attributes
  – Every level has members; the members of parent-child levels are in a one-to-many relation
  – The root level corresponds to the values of the dimension at the highest granularity
Examples: Hierarchies

\[
\text{year} \uparrow \
\text{semester} \uparrow \
\text{quarter} \uparrow \
\text{month} \uparrow \
\text{week} \uparrow \
\text{weekday} \uparrow \\
\text{day} \\
\text{customer} \uparrow \
\text{Address} \uparrow \
\text{City} \uparrow \
\text{Country} \uparrow \
\text{Age-category} \uparrow \
\text{Date-of-birth} \uparrow \
\]
Basic Concepts: Measure

- **Measure**: Numerical property of a fact; describes a quantitative aspect relevant for the analysis
  - Measures can be *aggregated*, grouping by the dimensions, using an *aggregation function* to form *secondary events*

Example: measure price; aggregation functions average, minimum, maximum
Notation: Dimensional Fact Model

Sale
- quantity
- unitPrice
- VAT-rate

customer
- country
- city
- age-category

product
- brand
- type

year
month
date
day-of-week

Notation: Dimensional Fact Model
Notation: Dimensional Fact Model

dimension with hierarchy

country  city  age-category  customer

fact

Sale

quantity  unitPrice  VAT-rate

measures

dimension with hierarchy

year  month  date

day-of-week

dimension with hierarchy

brand  type

dimension with hierarchy
Notation: Dimensional Fact Model

• There cannot be two sales for the same customer (C), date (D) and product (P)

• Customer can roll up to City and Age-category; City to Country
  \[ C \rightarrow \text{city} \quad C \rightarrow \text{age-group} \quad \text{city} \rightarrow \text{country} \]

• With every fact one quantity, unit price and VAT is associated
  \[ C,D,P \rightarrow \text{quantity}, \text{unitPrice}, \text{VAT} \]
Outline

• Dimensional fact model
  – Basic concepts
  – Extensions

• Roll-up lattice

• Special aggregation cases

• Properties of measures and aggregations

Chapter 5 of Golfarelli & Rizzi
Extensions

- country
- state
- city
- customer
- age-category
- day-of-week
- year
- month
- date
- quantity
- unitPrice
- VAT-rate
- product
- brand
- type
- promotion
- start
- discount

Incomplete/ragged hierarchy

Optional dimension

Descriptive attribute
Extension

• Optional level: is not specified for all members of the dimension

• Optional dimension: is not specified for all facts.
  – If promotion is missing, the other dimensions must be unique.

• Descriptive attribute: information that needs to be stored. But that is not suitable as a grouping attribute for aggregation
Example: Incomplete Hierarchies

country → state → city → customer

USA → Texas → Dallas → John

Belgium → Brussels → Peter George
Extensions

- Call duration cost number
- Tariff-plan
- Unit-cost
- country
- state
- city
- salesDistrict
- number
- calling
- called
- year
- month
- date
- day-of-week
- month
- date
- hour
- duration
- cost

Call

Diagram showing relationships between different elements such as country, state, city, salesDistrict, number, calling, called, year, month, date, day-of-week, unit-cost, year, hour, and duration cost.
Extensions

- Convergence
- Country
- Sales district
- State
- City
- Number
- Calling
- Called
- Unit cost
- Tariff plan
- Date
- Day of week
- Month
- Year
- Hour
- Convergence
- Shared hierarchy
- Cross-dimensional attribute
Example: Convergence

- Sometimes branches in a hierarchy merge again
Example: Recursive Hierarchy

salesman

Ahmed

John  Mary

Miranda

Pete

Patrick  Judith
In order to analyze the delays of their trains, a railway company decides to create a data warehouse in which they store all information relevant to the train delays. For every trip of a train that took place, the database should contain:

• The departure and destination station;
• The date of the trip;
• The planned departure and arrival times;
• The delay in minutes at arrival and at departure;
• The locomotive with which the trip was executed. Every locomotive has a unique number, a type, engine type (diesel or electricity), and total horsepower. There can be different locomotives of the same type. The type determines the engine type and the total horsepower.
• The driver. For the driver, his or her name, birth date, place of living, salary, and the types of trains he or she is allowed to conduct are stored as well.

Based on this data, the railway management would like to analyze, on a regular basis, the delays of the trains. In such analysis the train delays will typically be aggregated by time of the day, day of the week, by departure or destination station, or line (source-destination pair), and when systematic problems are detected on one or more lines, even an overview of the delays per driver on specific lines may be requested.
Outline

• Dimensional fact model
  – Basic concepts
  – Extensions
• Roll-up lattice
• Special aggregation cases
• Properties of measures and aggregations

Chapter 5 of Golfarelli & Rizzi
Roll-Up Lattice

• Dimensions and hierarchies define how data can be aggregated
  – A group-by set = set of incomparable dimensional attributes
  – Every primary event contributes to exactly one secondary event per group-by set

• Roll-up lattice = lattice containing all group-by sets, organized from general to specific
Roll-Up Lattice

Primary grouping set

Prod,Cust,Date
Roll-Up Lattice
Roll-Up Lattice

city

Customer

AgeCat

Month, Date

Sale

Product

Quantity

Unit Price

VAT Rate

Prod, Cust, Date

Prod, City, AgeCat, Date

Prod, Cust, Month

Cust, Date

City, AgeCat, Date

Cust, Month

City, Date

AgeCat, Date

City, AgeCat, Month
Roll-Up Lattice
Roll-Up Lattice

- city
- customer
- AgeCat
- month
date

<table>
<thead>
<tr>
<th>Sale</th>
</tr>
</thead>
<tbody>
<tr>
<td>quantity</td>
</tr>
<tr>
<td>unitPrice</td>
</tr>
<tr>
<td>VAT-rate</td>
</tr>
</tbody>
</table>

product

Prod, Cust, Date

Cust, Date

Prod, City, AgeCat, Date

Prod, Cust, Month

Prod, City, AgeCat, Date

Prod, City, Date

Prod, City, AgeCat, Month

Prod, City, Month

Prod, City, AgeCat

Prod, Month

Prod, AgeCat

Prod, City

Month

AgeCat

City

Prod
Outline

• Dimensional fact model
  – Basic concepts
  – Extensions
• Roll-up lattice
• Special aggregation cases
• Properties of measures and aggregations

Chapter 5 of Golfarelli & Rizzi
Optional Dimension

- Treat absence of a promotion as a “special value” (none; -)

<table>
<thead>
<tr>
<th>Month</th>
<th>Promotion</th>
<th>QTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>Discount 1</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Discount 2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>13</td>
</tr>
<tr>
<td>February</td>
<td>Discount 1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7</td>
</tr>
</tbody>
</table>
Incomplete Hierarchies

- What if level does not exist for some primary facts?
  - E.g., “state” for Brussels
# Incomplete Hierarchies

<table>
<thead>
<tr>
<th>Country</th>
<th>State</th>
<th>City</th>
<th>QTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Texas</td>
<td>Dallas</td>
<td>1</td>
</tr>
<tr>
<td>Belgium</td>
<td>-</td>
<td>Brussels</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>Antwerp</td>
<td>5</td>
</tr>
<tr>
<td>Germany</td>
<td>-</td>
<td>Berlin</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>Dresden</td>
<td>8</td>
</tr>
</tbody>
</table>

Roll-up to State

?
### Incomplete Hierarchies: Solution 1

<table>
<thead>
<tr>
<th>Country</th>
<th>State</th>
<th>City</th>
<th>QTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Texas</td>
<td>Dallas</td>
<td>1</td>
</tr>
<tr>
<td>Belgium</td>
<td>-</td>
<td>Brussels</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>Antwerp</td>
<td>5</td>
</tr>
<tr>
<td>Germany</td>
<td>-</td>
<td>Berlin</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>Dresden</td>
<td>8</td>
</tr>
</tbody>
</table>

**Roll-up to State**

<table>
<thead>
<tr>
<th>State</th>
<th>QTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texas</td>
<td>1</td>
</tr>
<tr>
<td>Other</td>
<td>21</td>
</tr>
</tbody>
</table>

**OR**

<table>
<thead>
<tr>
<th>Country</th>
<th>State</th>
<th>QTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Texas</td>
<td>1</td>
</tr>
<tr>
<td>Belgium</td>
<td>Other</td>
<td>10</td>
</tr>
<tr>
<td>Germany</td>
<td>Other</td>
<td>11</td>
</tr>
</tbody>
</table>
Balancing

Upward balancing

country
state
city

USA
Texas
Dallas

Belgium
Brussels
Antwerp

Germany
Berlin
Dresden

Downward balancing

country
state
city

USA
Texas
Dallas

Belgium
Brussels
Antwerp

Germany
Berlin
Dresden
# Upward Balancing

<table>
<thead>
<tr>
<th>Country</th>
<th>State</th>
<th>City</th>
<th>QTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Texas</td>
<td>Dallas</td>
<td>1</td>
</tr>
<tr>
<td>Belgium</td>
<td>Belgium</td>
<td>Brussels</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Antwerp</td>
<td>5</td>
</tr>
<tr>
<td>Germany</td>
<td>Germany</td>
<td>Berlin</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dresden</td>
<td>8</td>
</tr>
</tbody>
</table>

Roll-up to State

<table>
<thead>
<tr>
<th>Country</th>
<th>State</th>
<th>QTY</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>Texas</td>
<td>1</td>
</tr>
<tr>
<td>Belgium</td>
<td>Belgium</td>
<td>10</td>
</tr>
<tr>
<td>Germany</td>
<td>Germany</td>
<td>11</td>
</tr>
</tbody>
</table>
Multiple Arcs

- Roll-up by author can be misleading
  - Sales of same book will be counted multiple times
- Solution: add edge weights
Multiple Arcs: without weights

<table>
<thead>
<tr>
<th>Book</th>
<th>Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>A2</td>
</tr>
<tr>
<td>B2</td>
<td>A1</td>
</tr>
<tr>
<td></td>
<td>A3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Author</th>
<th>Count()</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>2</td>
</tr>
<tr>
<td>A2</td>
<td>1</td>
</tr>
<tr>
<td>A3</td>
<td>1</td>
</tr>
</tbody>
</table>

Total: 4
# Multiple Arcs: with weights

## Table

<table>
<thead>
<tr>
<th>Book</th>
<th>Author</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>B1</td>
<td>A1</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>A2</td>
<td>0.5</td>
</tr>
<tr>
<td>B2</td>
<td>A1</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>A3</td>
<td>0.3</td>
</tr>
</tbody>
</table>

## Weighted Count

<table>
<thead>
<tr>
<th>Author</th>
<th>Weighted Count()</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>1.2</td>
</tr>
<tr>
<td>A2</td>
<td>0.5</td>
</tr>
<tr>
<td>A3</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Total: 2
Cross-Dimensional Attributes

• VAT results in extra levels in roll-up
  – VAT; VAT,country; VAT,type
  – VAT,country,type or VAT,country,product (redundant)
Cross-Dimensional Attributes

P,C
B,T,C
P,V
B,T,V
T,C
B,C,V
B,T,V
P
B,
C,V
T,V
B,
C
B,
V
B,
T
C
V
T
B
{}

Sale
quantity

country

brand

product
type

VAT
Measureless Schemas

• Some schemas do not have measures, but it could still be interesting to aggregate
  – COUNT; AND; OR
# Measureless Schemas

<table>
<thead>
<tr>
<th>Year</th>
<th>Name</th>
<th>BPM</th>
<th>DW</th>
<th>ADB</th>
<th>DBSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>John</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Mary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Pete</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Patrick</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td>Patrick</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Jane</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Pete</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>
## Measureless Schemas

### COUNT

<table>
<thead>
<tr>
<th>Year</th>
<th>BPM</th>
<th>DW</th>
<th>ADB</th>
<th>DBSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2013</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

### OR

<table>
<thead>
<tr>
<th>Name</th>
<th>BPM</th>
<th>DW</th>
<th>ADB</th>
<th>DBSA</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Mary</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pete</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Patrick</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Jane</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>
Measureless Schemas

Combination of OR on Year and AND on Course

<table>
<thead>
<tr>
<th></th>
<th>AND</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>1</td>
</tr>
<tr>
<td>Mary</td>
<td>0</td>
</tr>
<tr>
<td>Pete</td>
<td>0</td>
</tr>
<tr>
<td>Patrick</td>
<td>1</td>
</tr>
<tr>
<td>Jane</td>
<td>0</td>
</tr>
</tbody>
</table>
Outline

• Dimensional fact model
  – Basic concepts
  – Extensions

• Roll-up lattice

• Special aggregation cases

• Properties of measures and aggregations

Chapter 5 of Golfarelli & Rizzi
Non-Additive Measures

• A measure is *non-additive* over a dimension if you cannot use the SUM operator to aggregate its values over that dimension

• *Non-aggregable*: If no aggregation operator can be used

Example of non-additive measures:
- Stock level over time
- Unit price over time or customer
Non-Additive Measures

- Complement model with additivity matrix

<table>
<thead>
<tr>
<th></th>
<th>date</th>
<th>warehouse</th>
<th>product</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>AVG, MIN, MAX</td>
<td>SUM, AVG, MIN, MAX</td>
<td>SUM, AVG, MIN, MAX</td>
</tr>
<tr>
<td>IncomingQTY</td>
<td>SUM, AVG, MIN, MAX</td>
<td>SUM, AVG, MIN, MAX</td>
<td>SUM, AVG, MIN, MAX</td>
</tr>
</tbody>
</table>
Non-Additive Measures

• Or, indicate directly in the schema
Distributive Operators

• SUM, MIN, MAX are distributive aggregation operators
  – \( \text{SUM}(A,B,C,D) = \text{SUM}(\text{SUM}(A,B),\text{SUM}(C,D)) = \text{SUM}(\text{SUM}(A,B,C),D) = \ldots \)

• For distributive operators it holds:
  – If \( G_1 \) above \( G_2 \) in roll-up lattice, then group by \( G_2 \) can be computed directly from group by \( G_1 \)
Example: Distributive Operator

<table>
<thead>
<tr>
<th>Continent</th>
<th>Country</th>
<th>City</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>Belgium</td>
<td>Brussels</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Antwerp</td>
<td>3</td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td>Berlin</td>
<td>2</td>
</tr>
<tr>
<td>North-America</td>
<td>USA</td>
<td>Chicago</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tampa</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Continent</th>
<th>SUM(Amount)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>10</td>
</tr>
<tr>
<td>North-America</td>
<td>9</td>
</tr>
</tbody>
</table>

Sum(Amount) = 19
**Example: Non-Distributive Operator**

<table>
<thead>
<tr>
<th>Continent</th>
<th>Country</th>
<th>City</th>
<th>Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>Belgium</td>
<td>Brussels</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Antwerp</td>
<td>3</td>
</tr>
<tr>
<td>Germany</td>
<td></td>
<td>Berlin</td>
<td>2</td>
</tr>
<tr>
<td>North-America</td>
<td>USA</td>
<td>Chicago</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tampa</td>
<td>8</td>
</tr>
</tbody>
</table>

**AVG(Amount)**

<table>
<thead>
<tr>
<th>Continent</th>
<th>AVG(Amount)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Europe</td>
<td>3.33</td>
</tr>
<tr>
<td>North-America</td>
<td>4.5</td>
</tr>
</tbody>
</table>
Algebraic Operator

• With some additional information some non-distributive aggregation operators can still be calculated from partial aggregates
  – $\text{AVG}(A,B,C,D) = \frac{\text{SUM}(A,B,C,D)}{\text{COUNT}(A,B,C,D)}$
  – $\text{VAR}(A,B,C,D) = \text{AVG}(A^2,B^2,C^2,D^2) - \text{AVG}(A,B,C,D)^2$

• Such operators are called “algebraic”
  – With support measures we have more efficient aggregation
  – Impacts the logical design
Holistic Operator

• Operator that is distributive nor algebraic
  – Median, mode

• For these operators, the only way to compute the secondary events is from the primary facts
  – consequences for efficiency