Data Warehousing
Dimensional Fact Model

Esteban Zimányi
ezimanyi@ulb.ac.be
Slides by Toon Calders

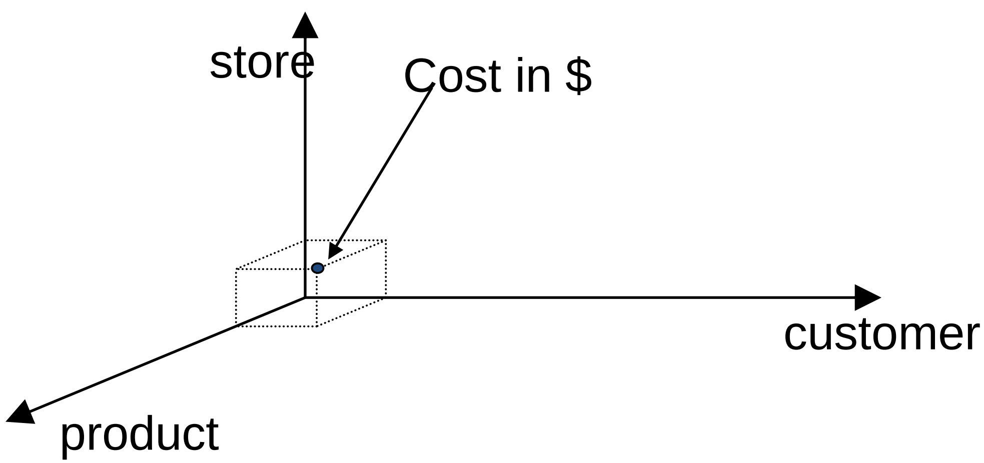
ECOLE
POLYTECHNIQUE
DE BRUXELLES

What have we seen last time?

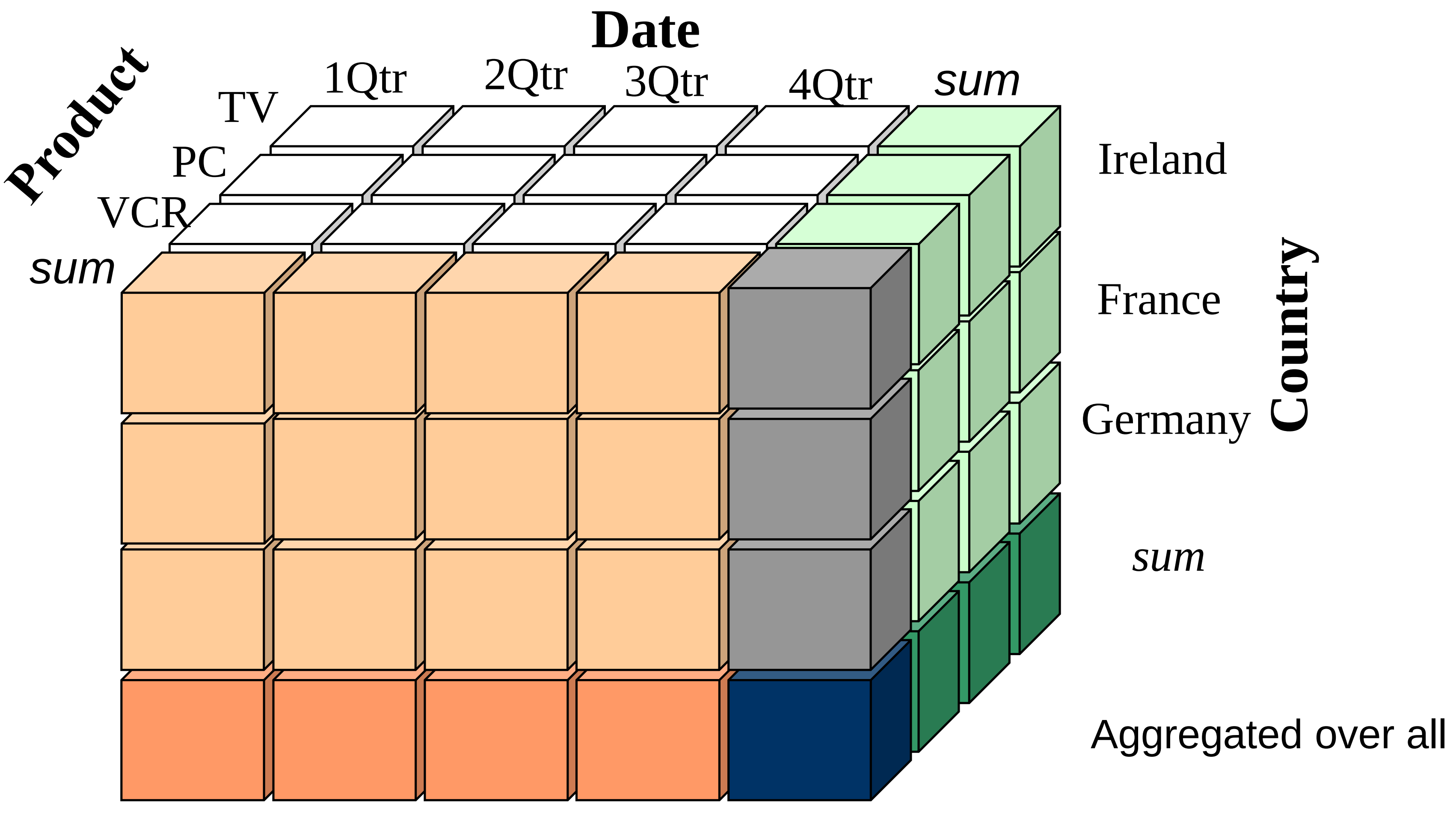
- Evaluate the sales of products
  - Product cost in $ — measure
  - Customer: ID, city, state, country,
  - Store: chain, size, location,
  - Product: brand, type, ...
  - ...

Dim.

hierarchies



What have we seen last time?



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

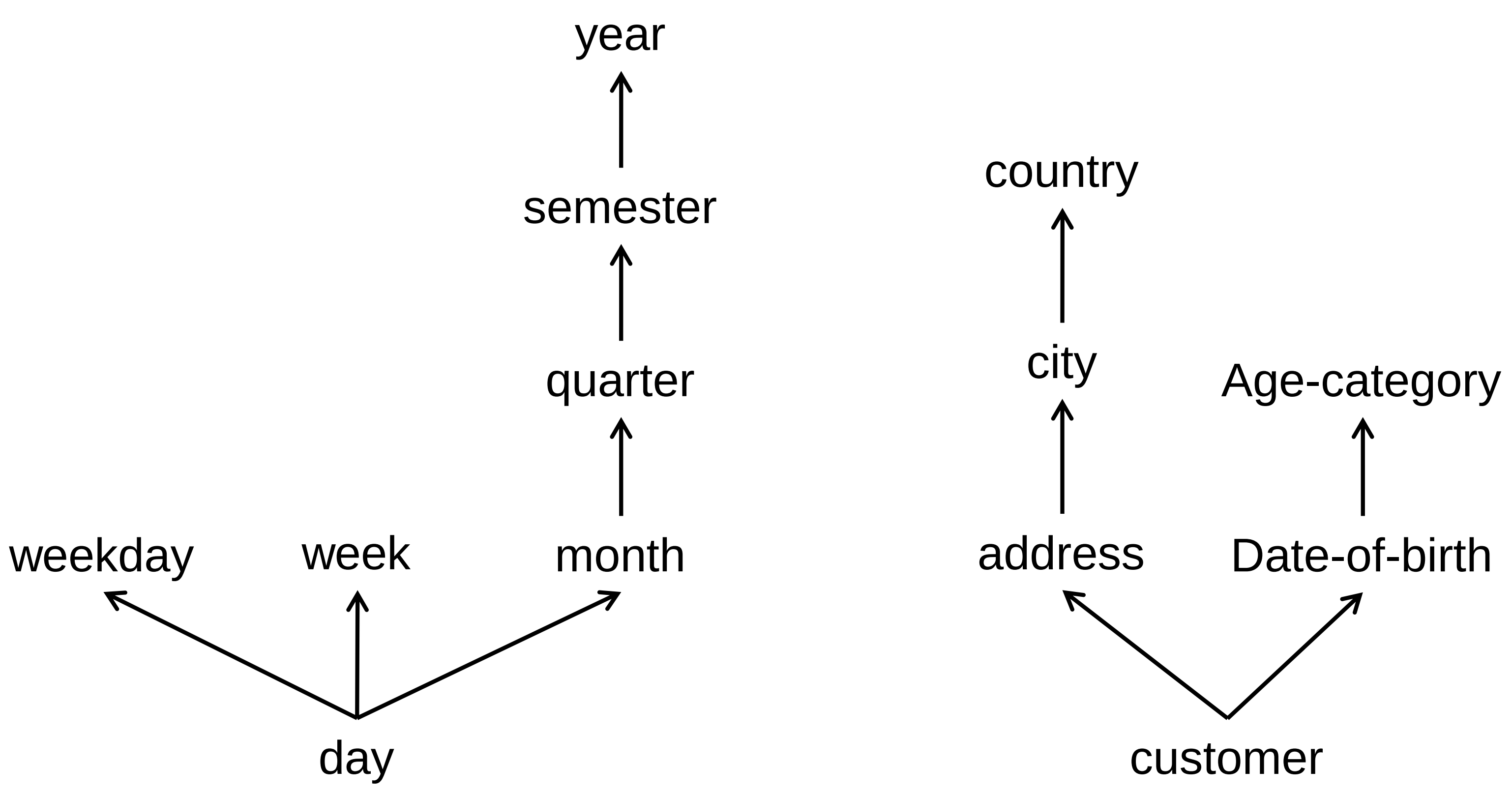
Examples: Dimension

- 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

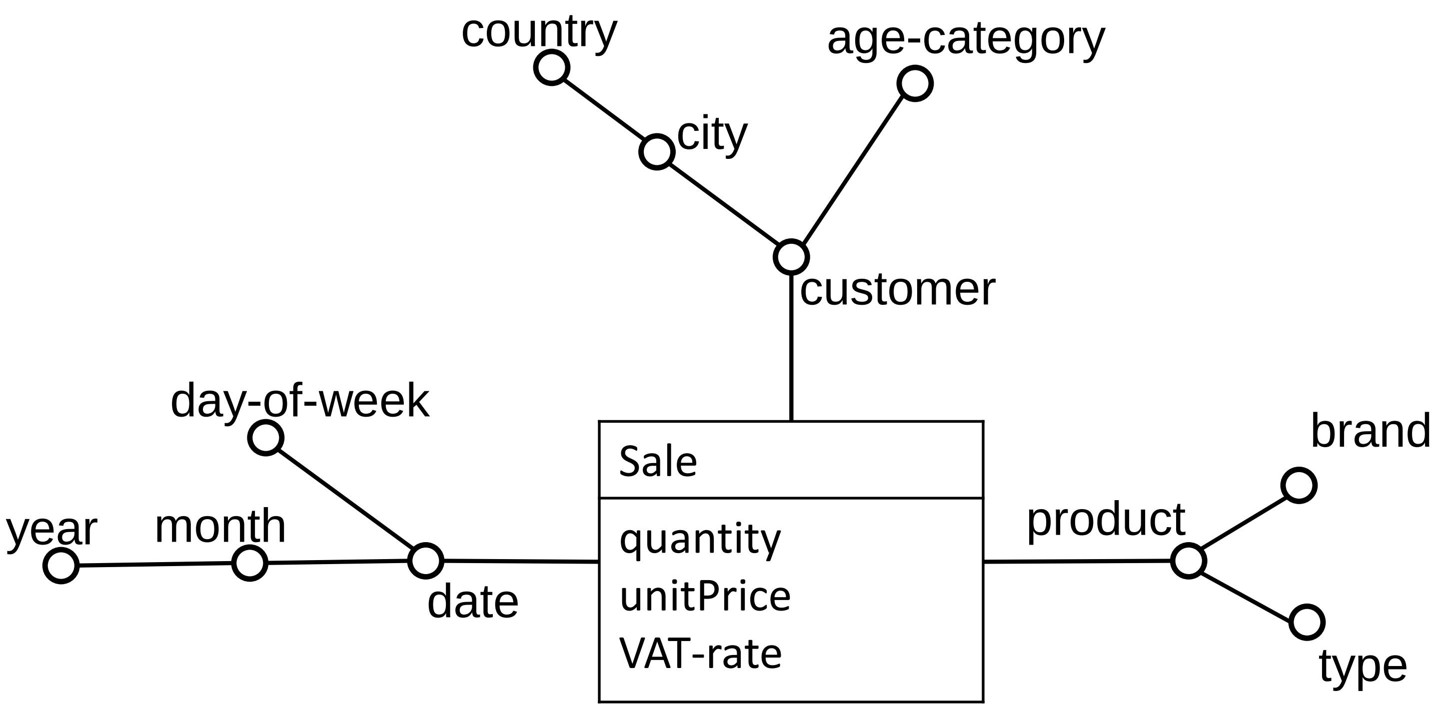


Basic Concepts: Measure

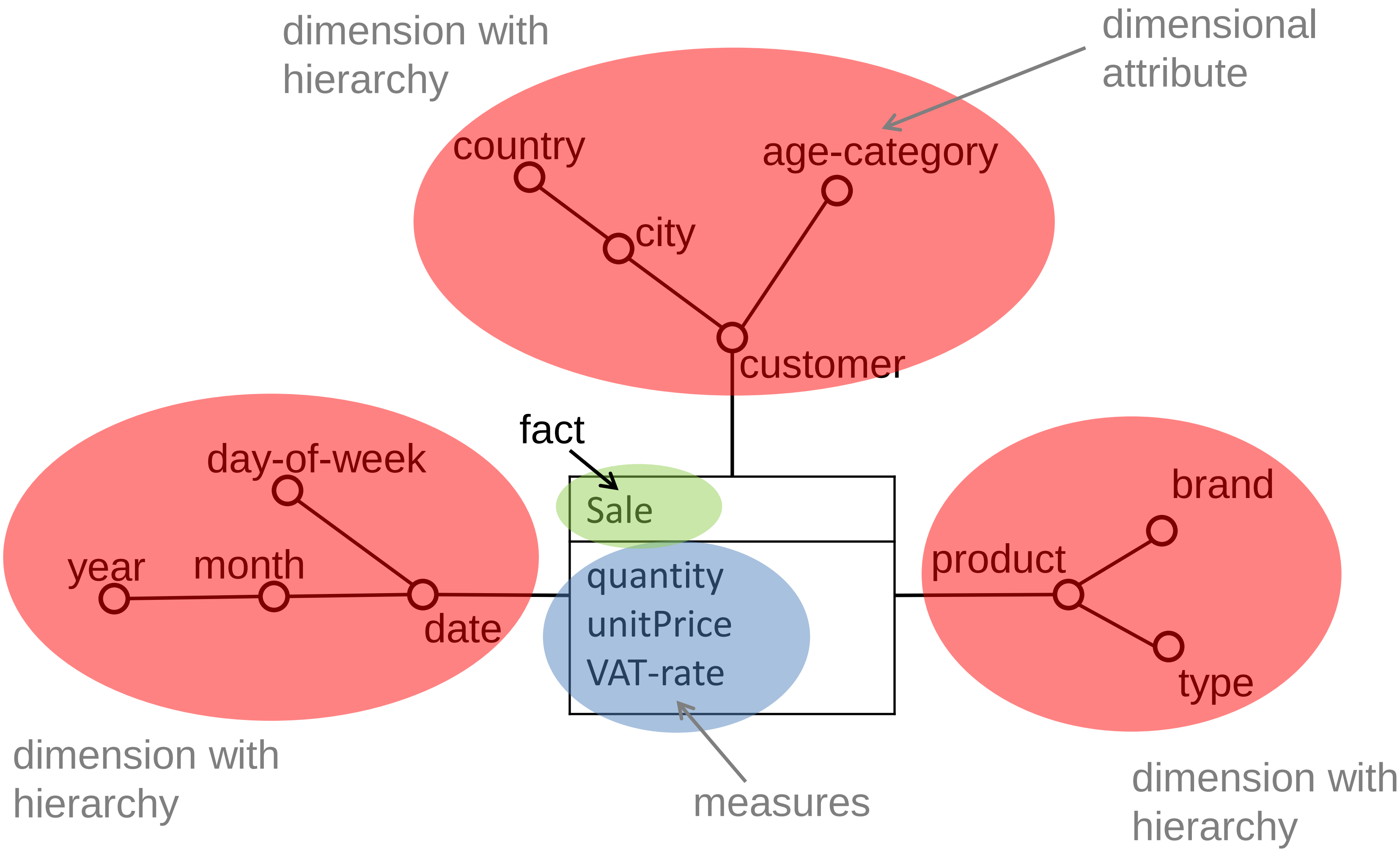
- *Measure*: Numerical property of a fact; describes a quantitative aspect relevant for the analysis
  - Measures can be *aggregated*, grouping by the dimensions, using an *aggregation function* to form *secondary events*

Example: measure price; aggregation functions average, minimum, maximum

Notation: Dimensional Fact Model



Notation: Dimensional Fact Model



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 → city   C → age-group   city → country

- With every fact one quantity, unit price and VAT is associated

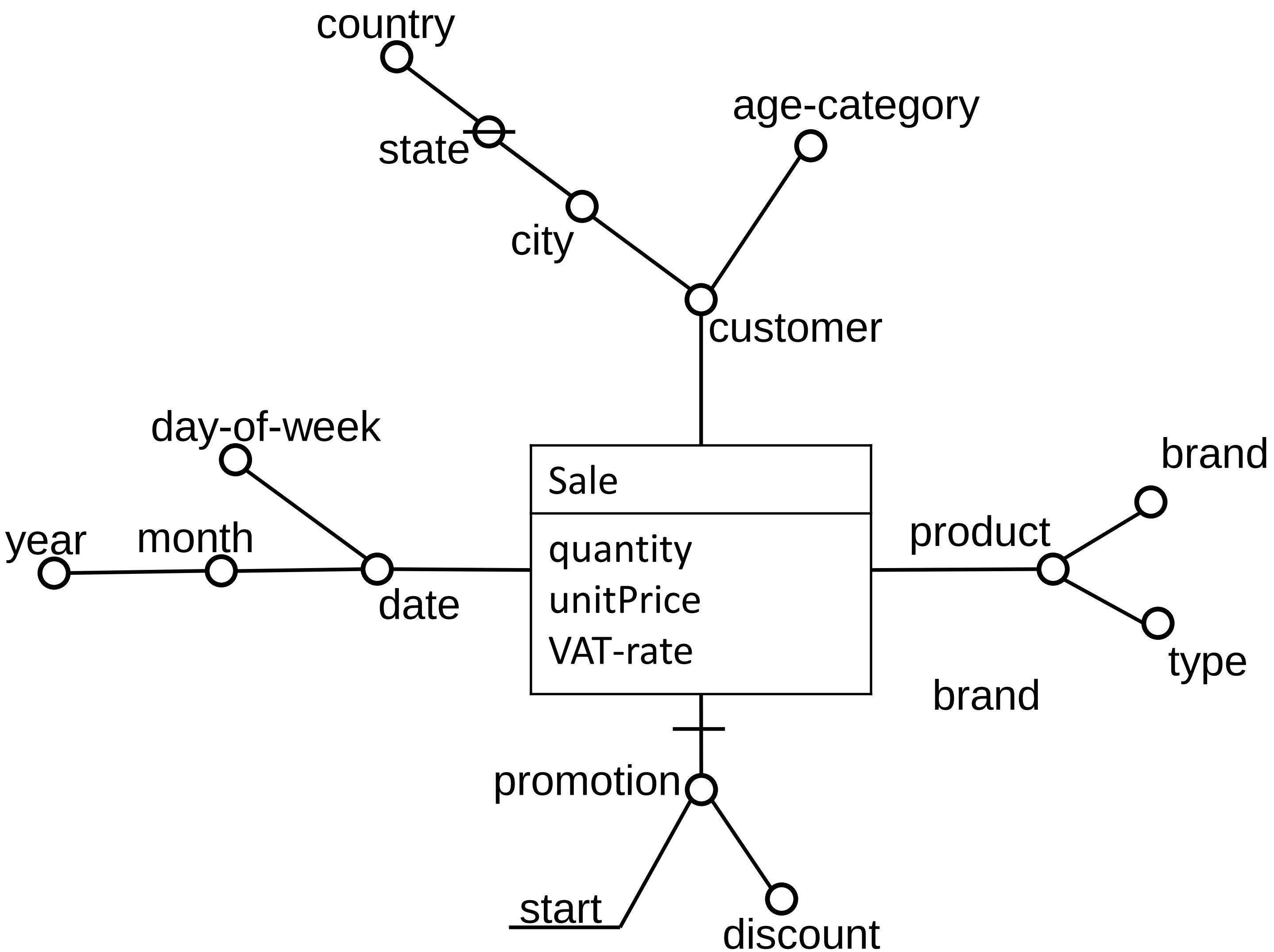
  C,D,P → quantity, unitPrice, VAT

Outline

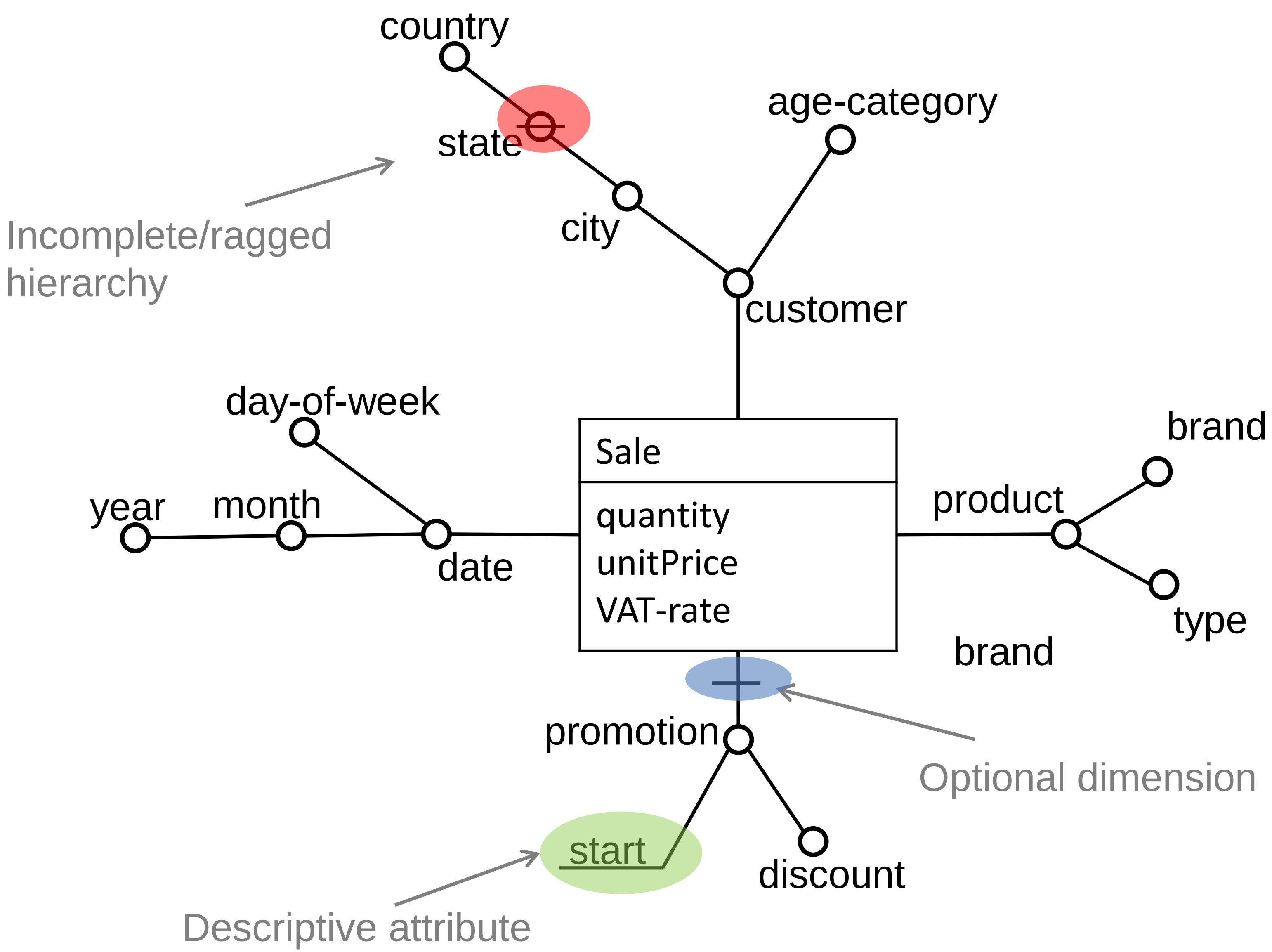
- Dimensional fact model
  - Basic concepts
  - Extensions
- Roll-up lattice
- Special aggregation cases
- Properties of measures and aggregations

Chapter 5 of Golfarelli & Rizzi

Extensions


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

Extensions

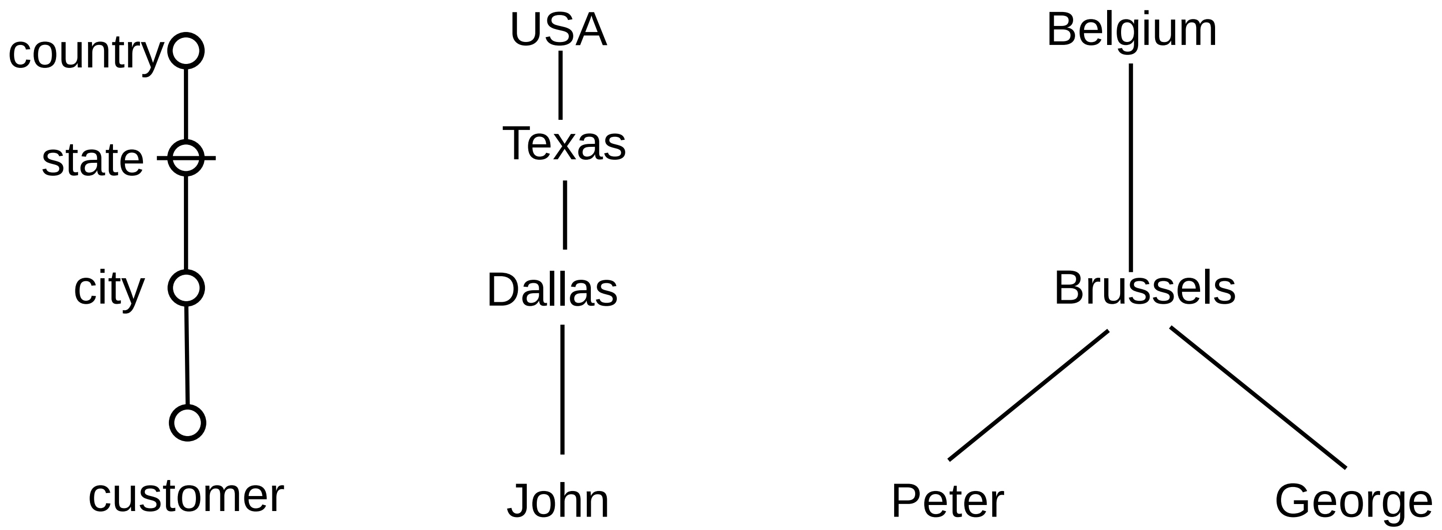


country
state
age-category
city
customer
Incomplete/ragged
hierarchy
day-of-week
Sale
quantity
unitPrice
VAT-rate
brand
product
year
month
date
type
brand
promotion
Optional dimension
start
discount
Descriptive attribute

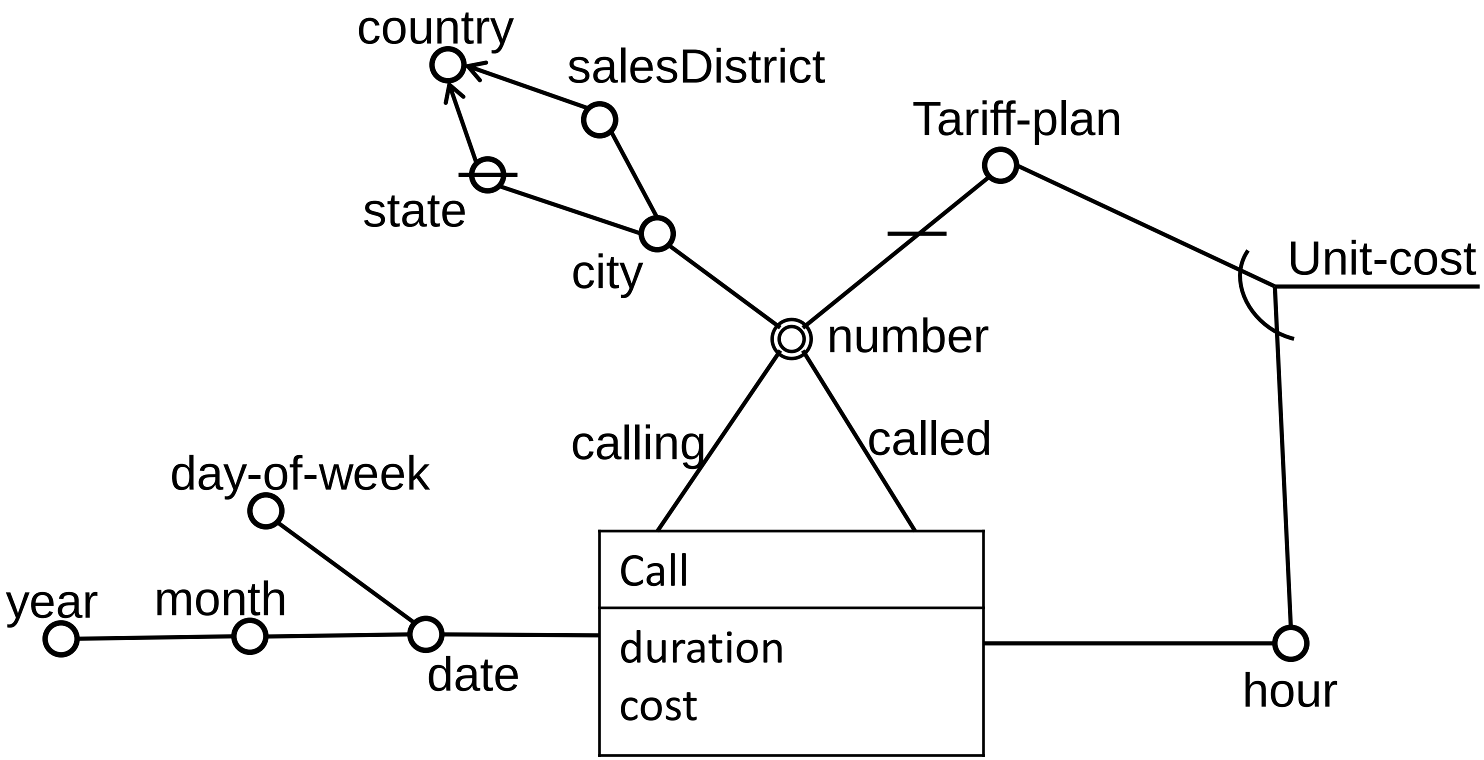
Extension

- 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

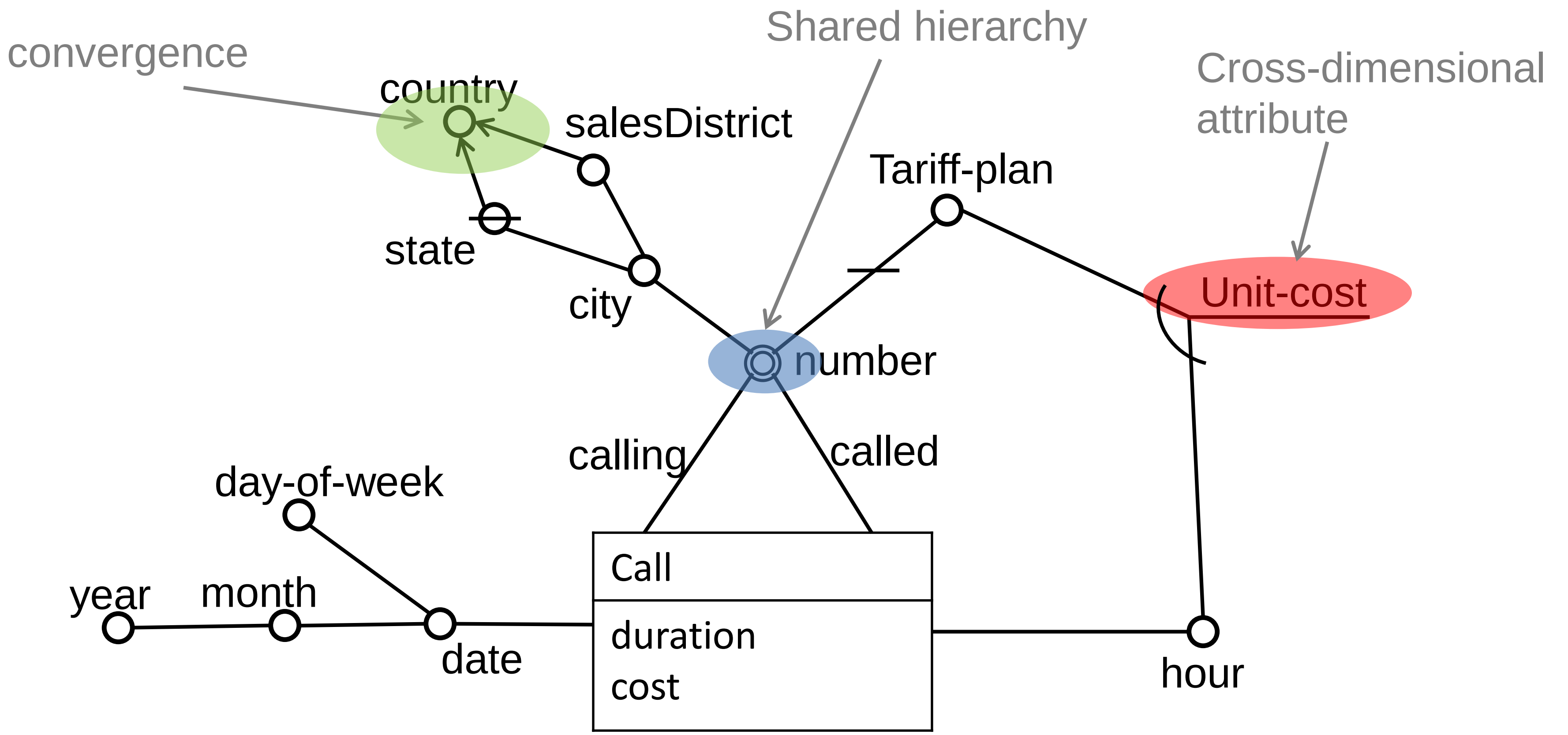


Extensions



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

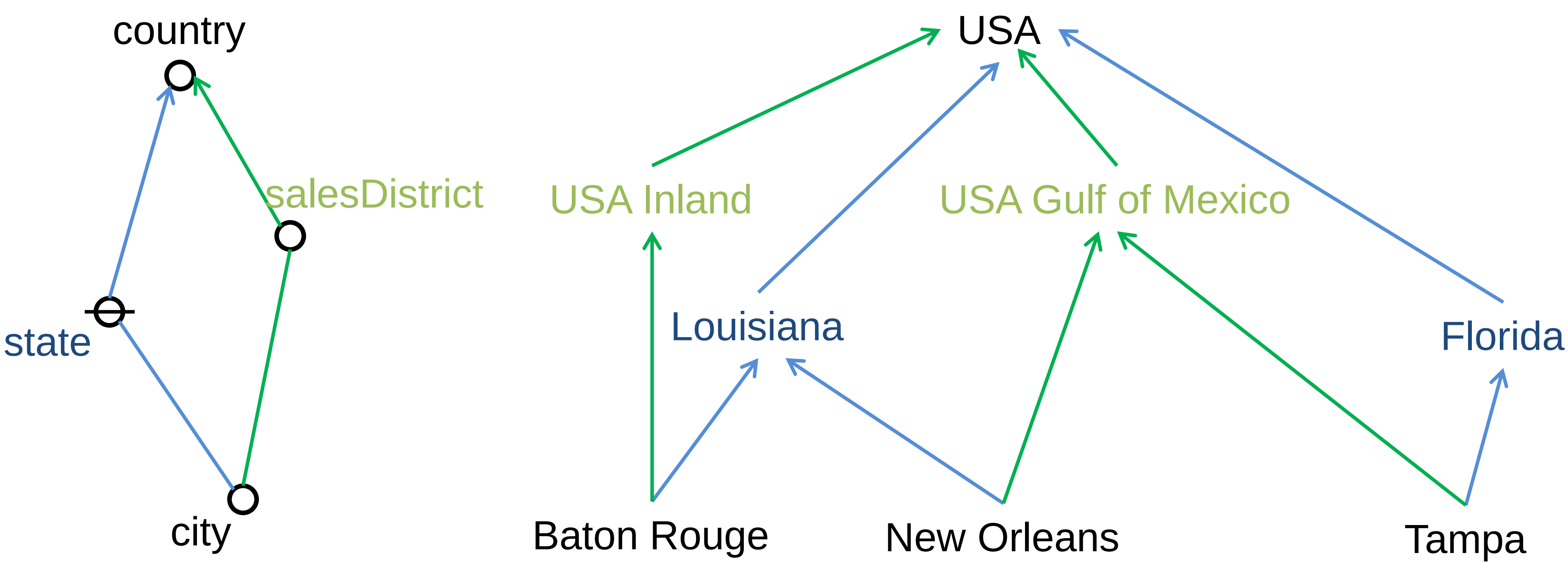
Extensions



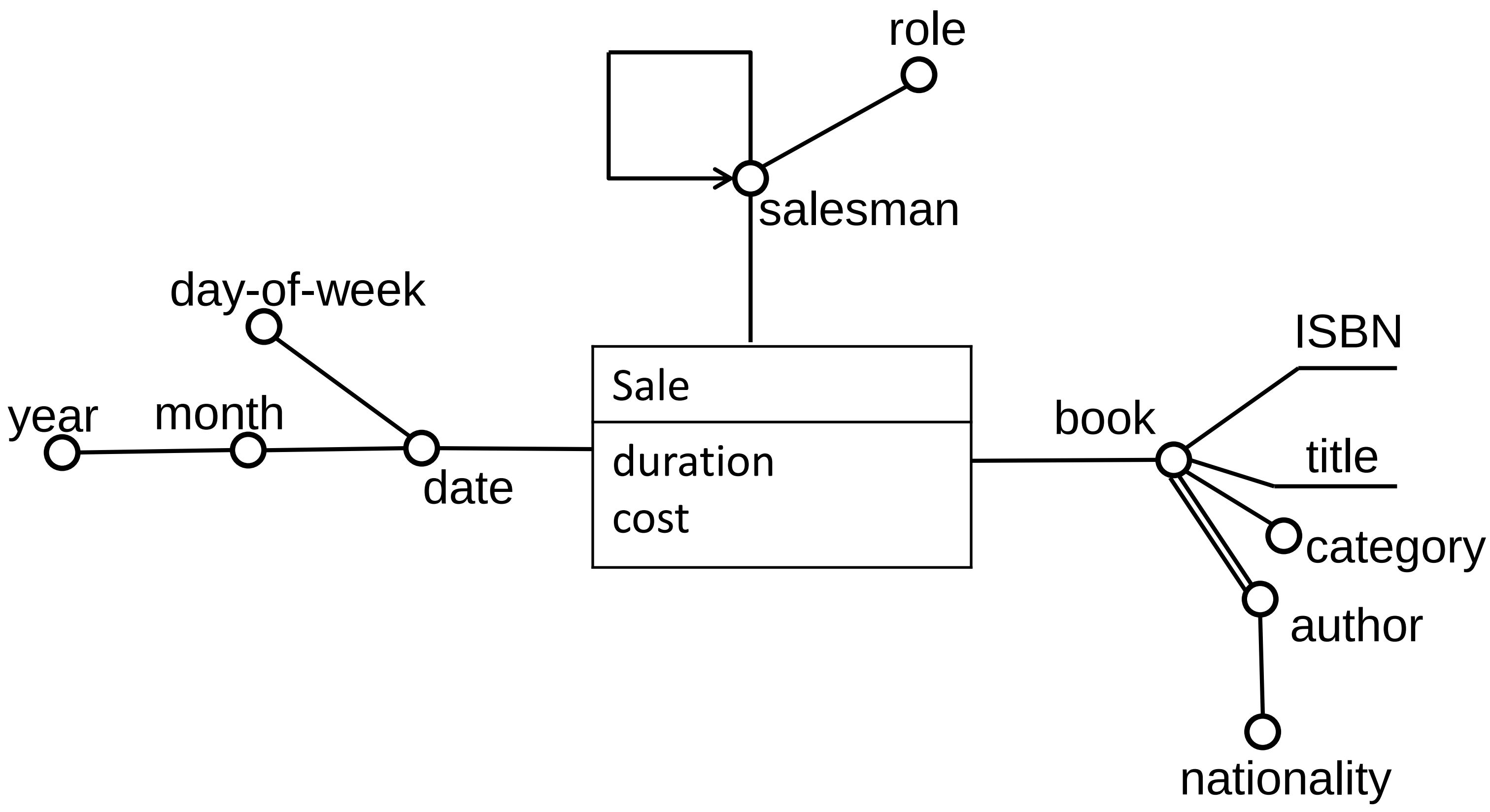
convergence
Shared hierarchy
Cross-dimensional
attribute
country
salesDistrict
Tariff-plan
state
city
Unit-cost
number
calling
called
day-of-week
Call
duration
cost
year
month
date
hour

Example: Convergence

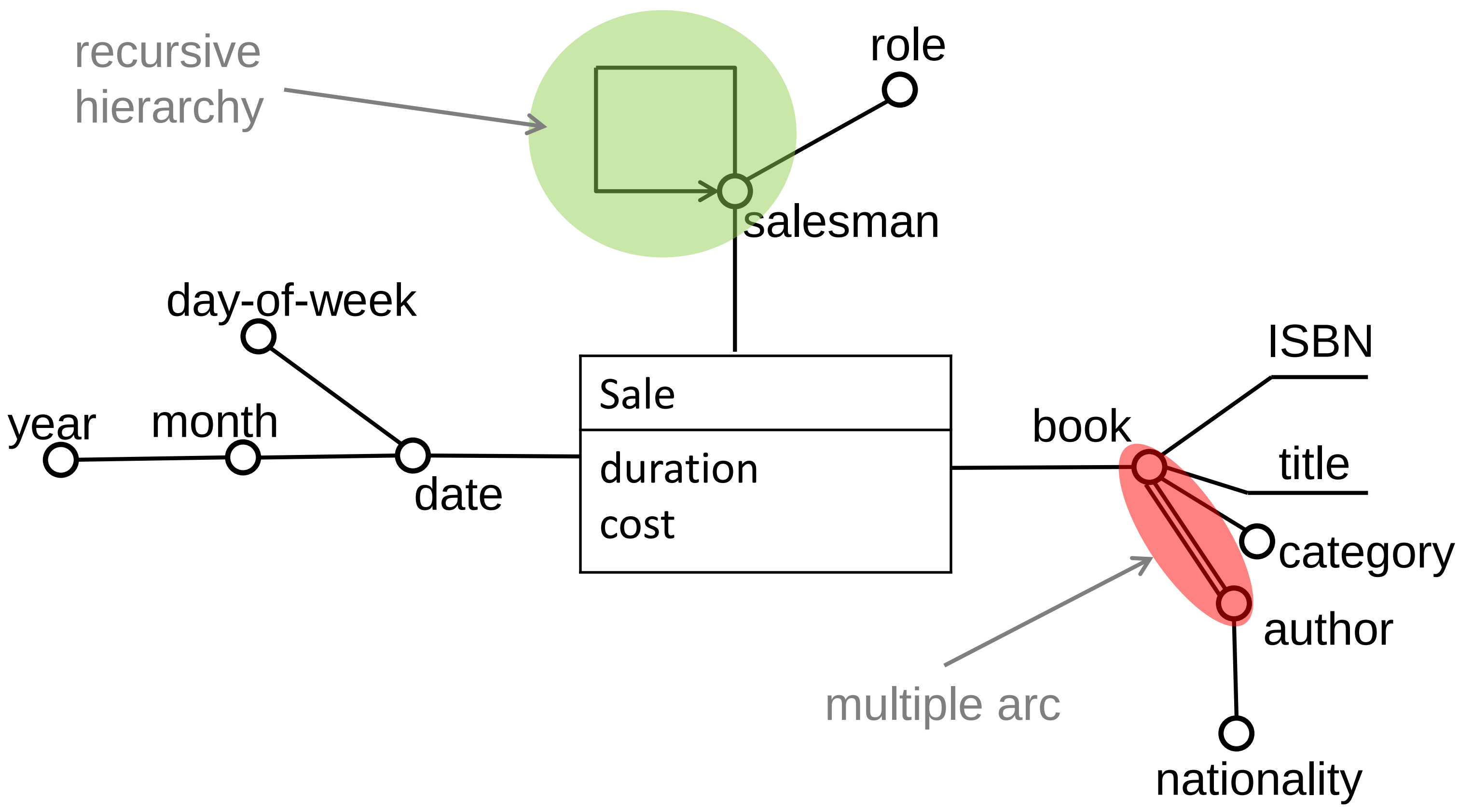
- Sometimes branches in a hierarchy merge again



Extensions

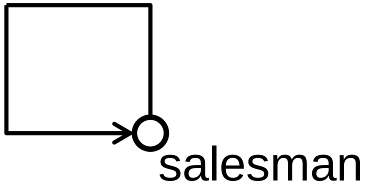

role
salesman
day-of-week
year
month
date
Sale
duration
cost
book
ISBN
title
category
author
nationality

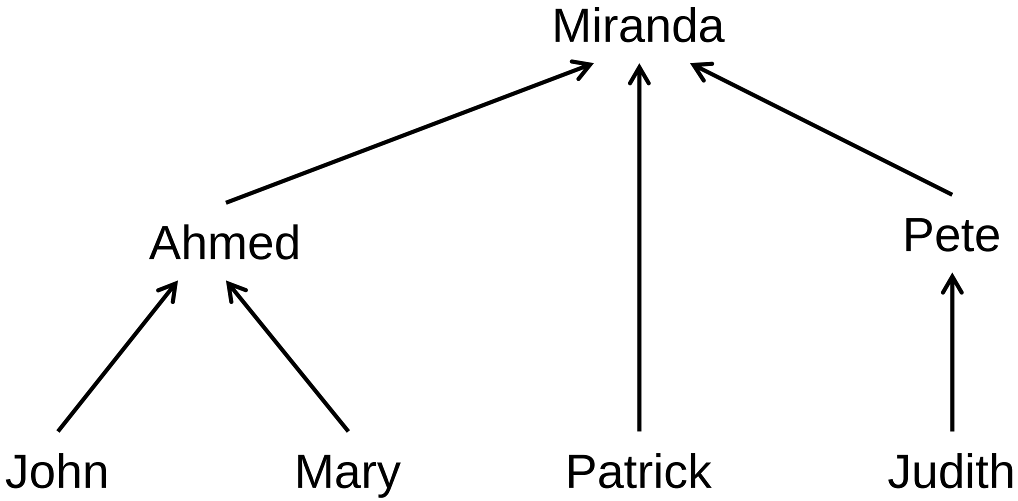
Extensions



recursive
hierarchy
role
salesman
day-of-week
ISBN
year
month
Sale
duration
cost
book
title
date
category
author
multiple arc
nationality

Example: Recursive Hierarchy





Exercise

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

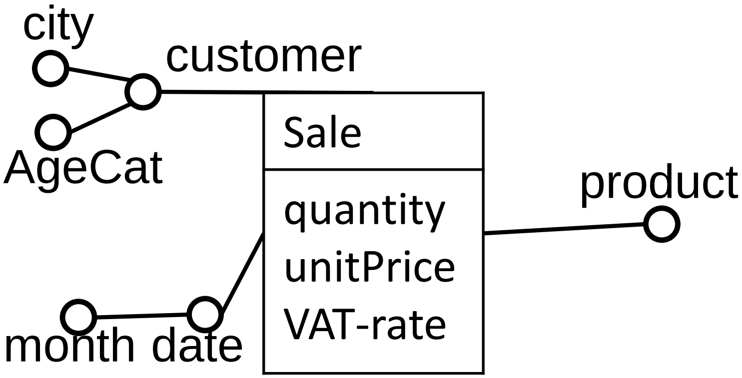
- Dimensional fact model
  - Basic concepts
  - Extensions
- Roll-up lattice
- Special aggregation cases
- Properties of measures and aggregations

Chapter 5 of Golfarelli & Rizzi

Roll-Up Lattice

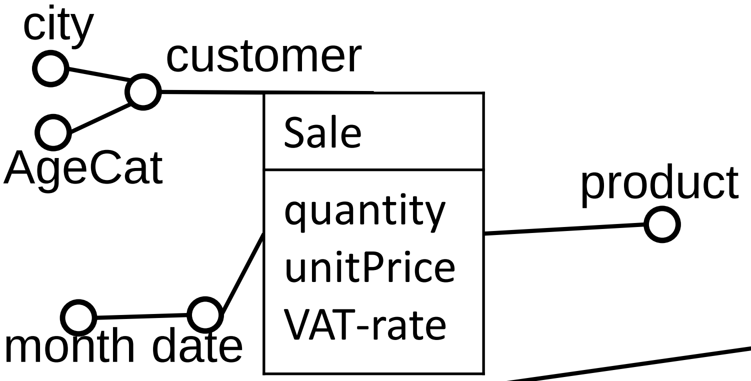
- Dimensions and hierarchies define how data can be aggregated
  - A *group-by set* = set of incomparable dimensional attributes
  - Every primary event contributes to exactly one secondary event per group-by set
- Roll-up lattice = lattice containing all group-by sets, organized from general to specific

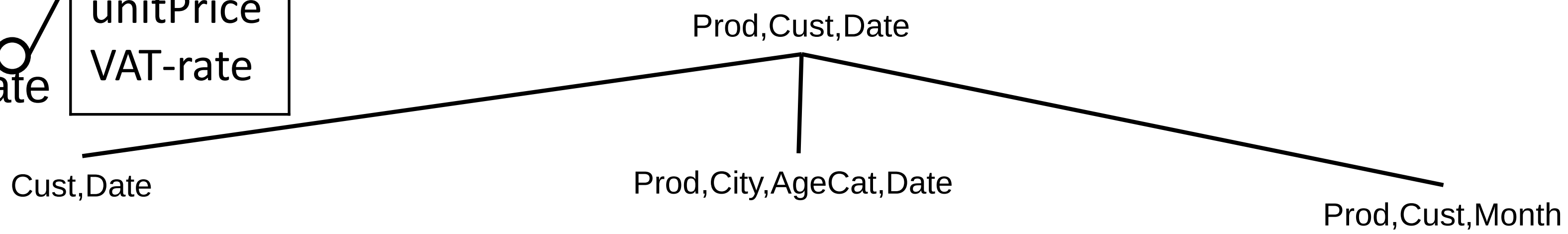
Roll-Up Lattice



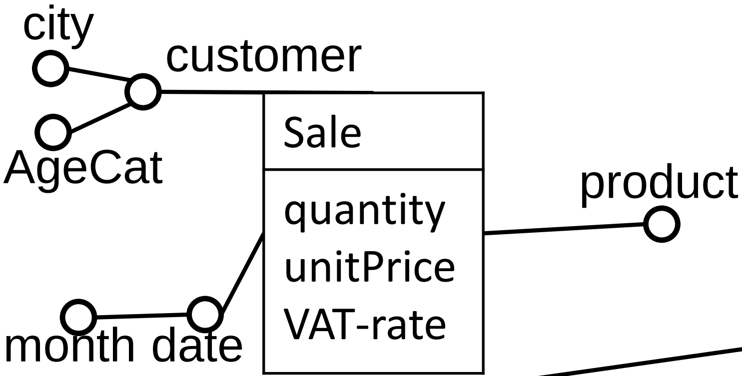
Prod,Cust,Date ← Primary grouping set

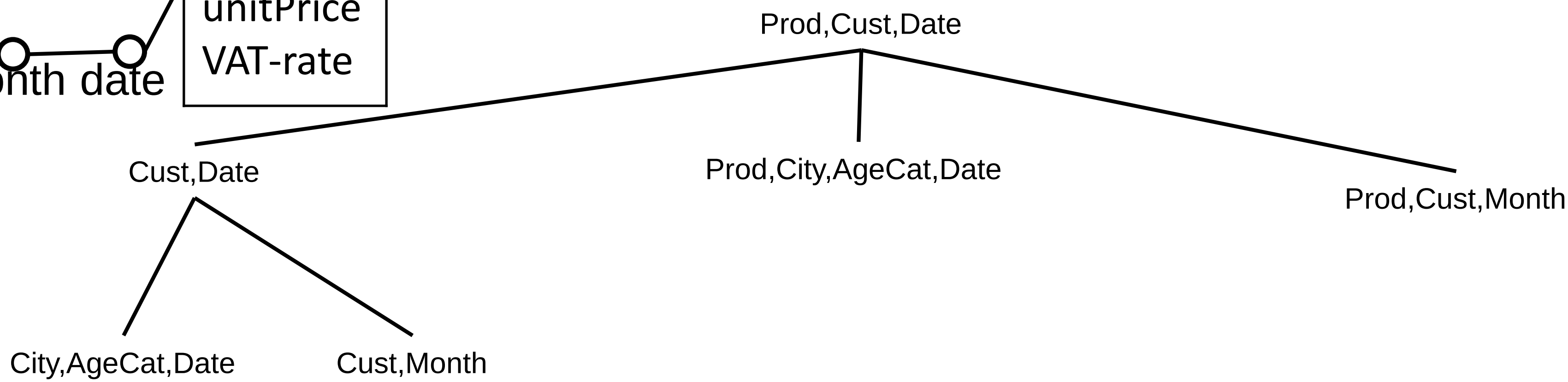
Roll-Up Lattice



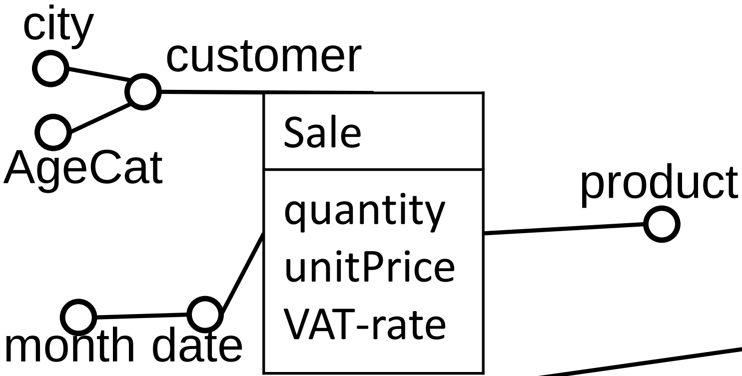


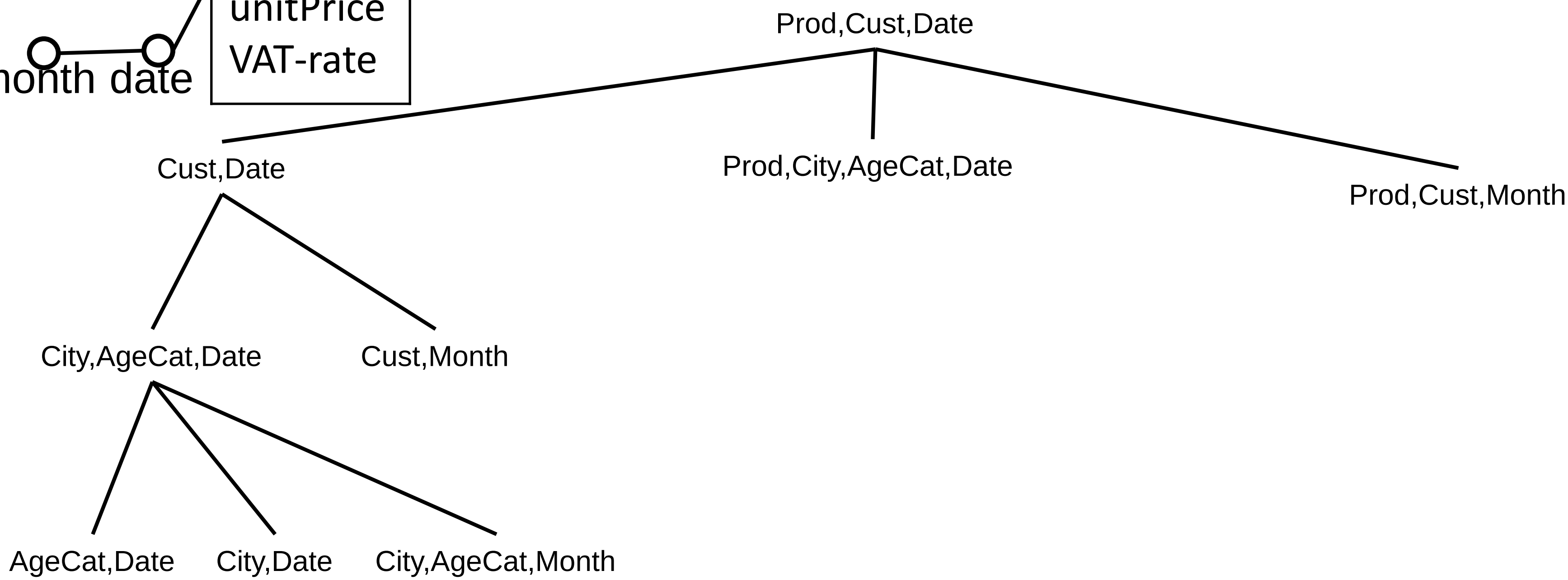
Roll-Up Lattice



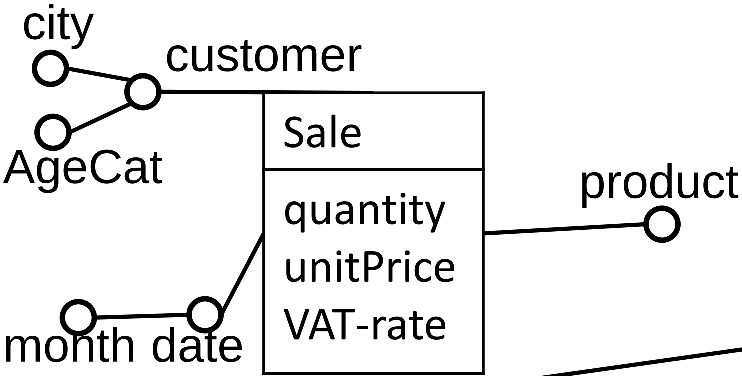


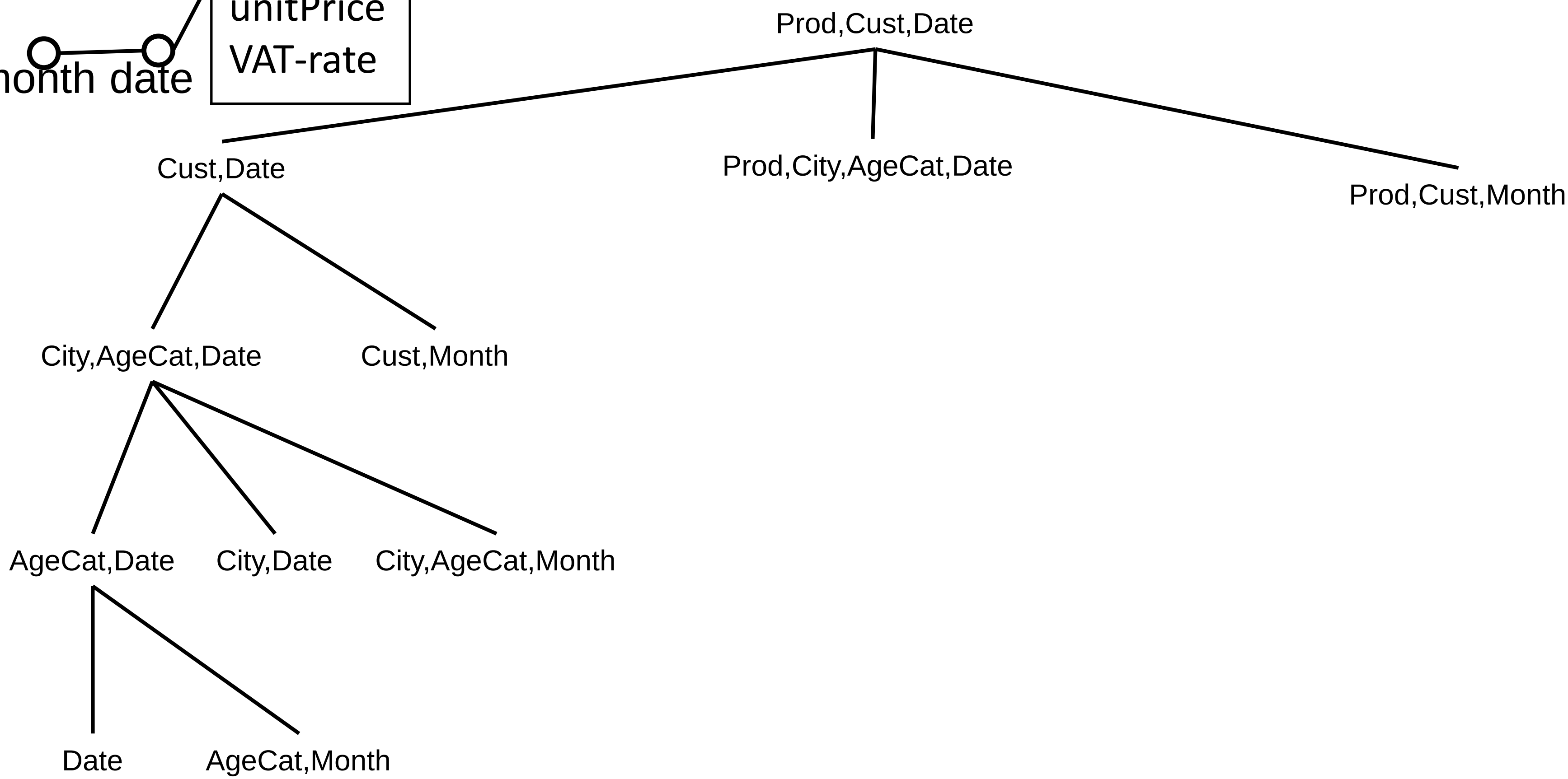
Roll-Up Lattice



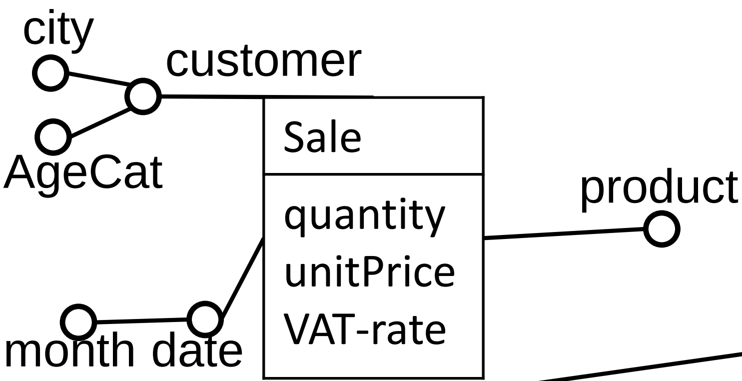


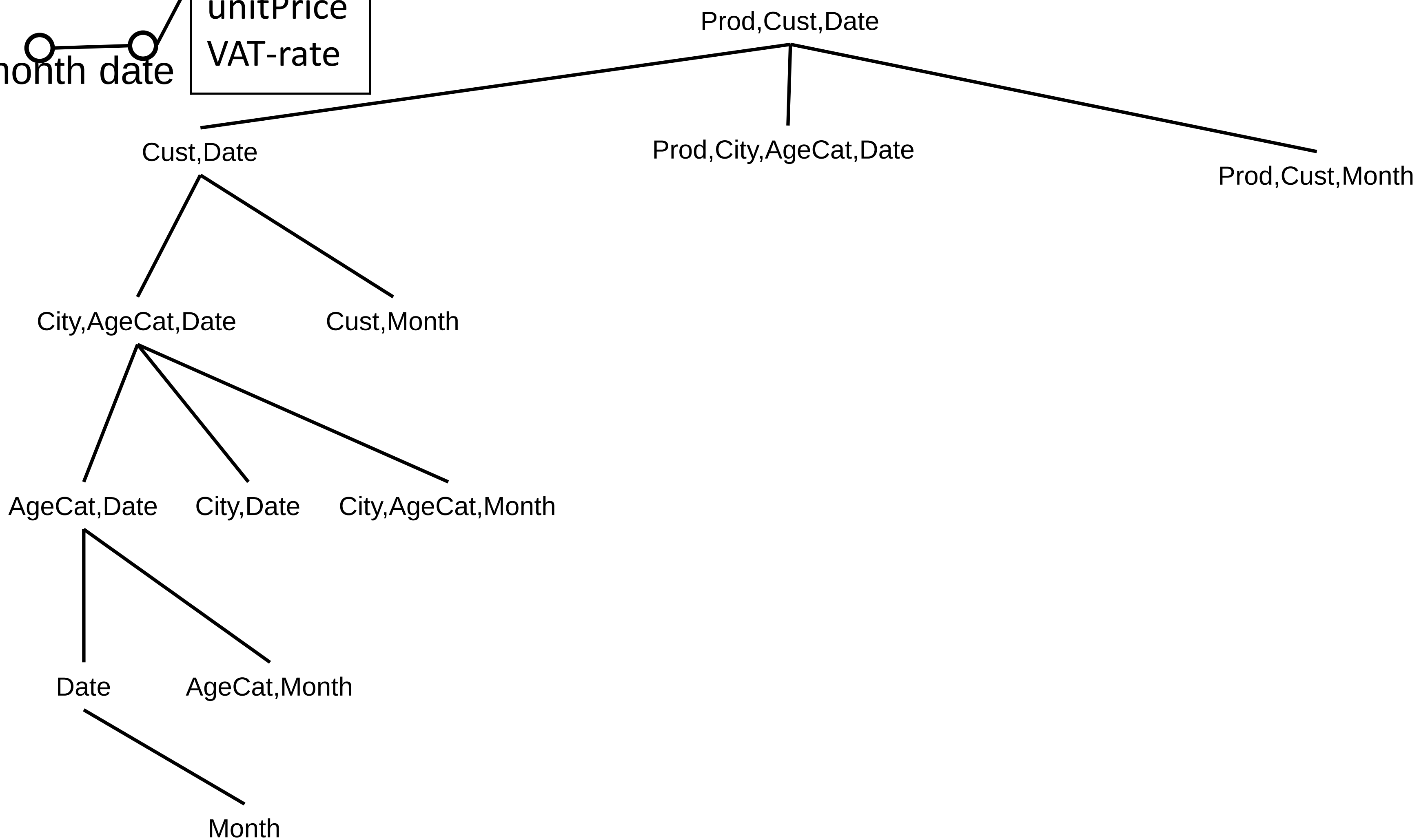
Roll-Up Lattice



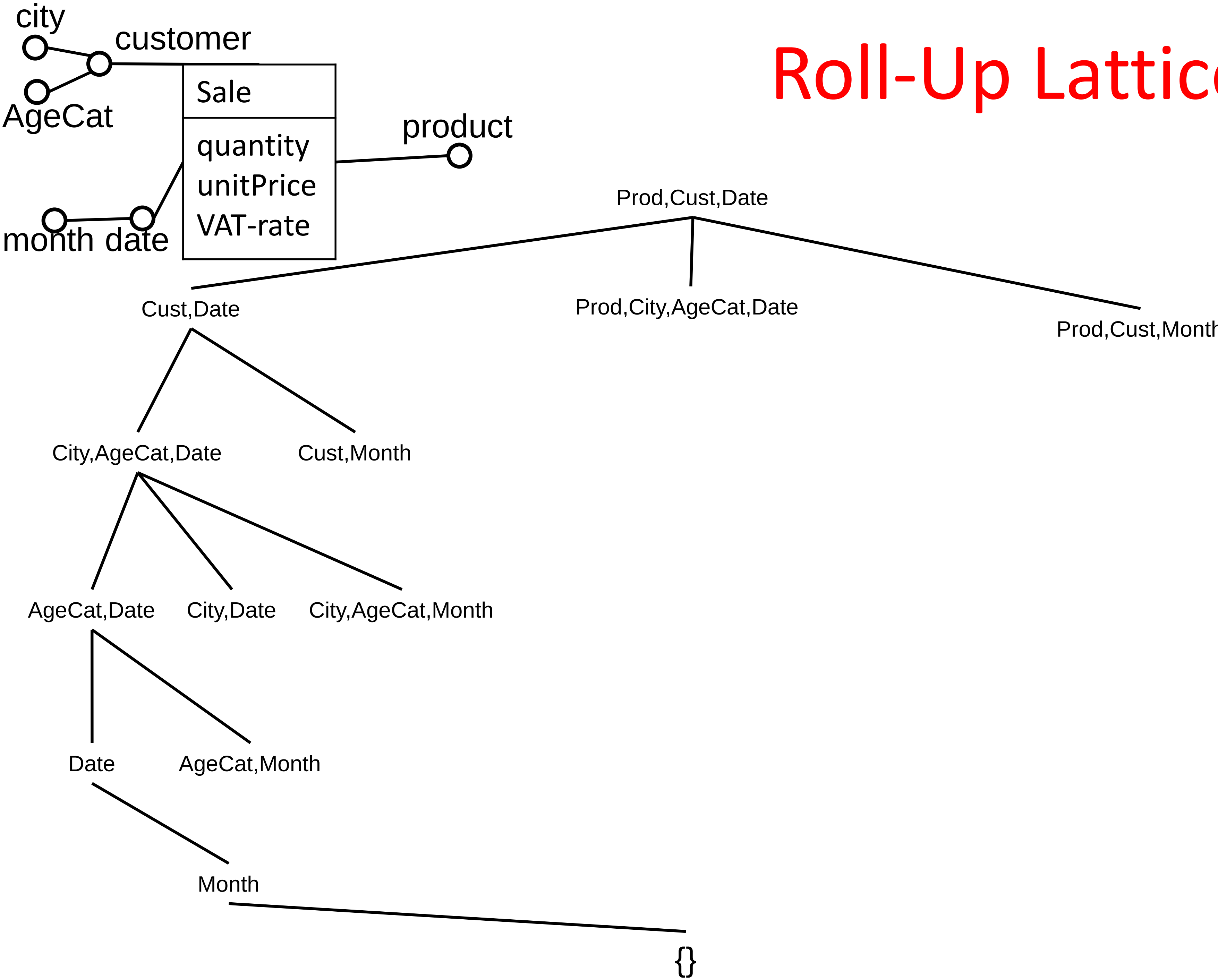


Roll-Up Lattice

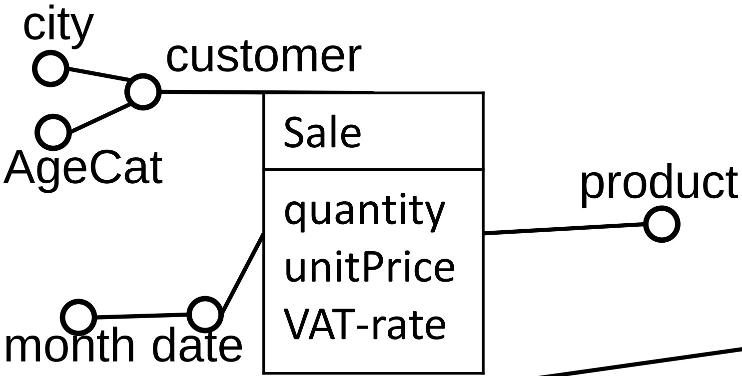


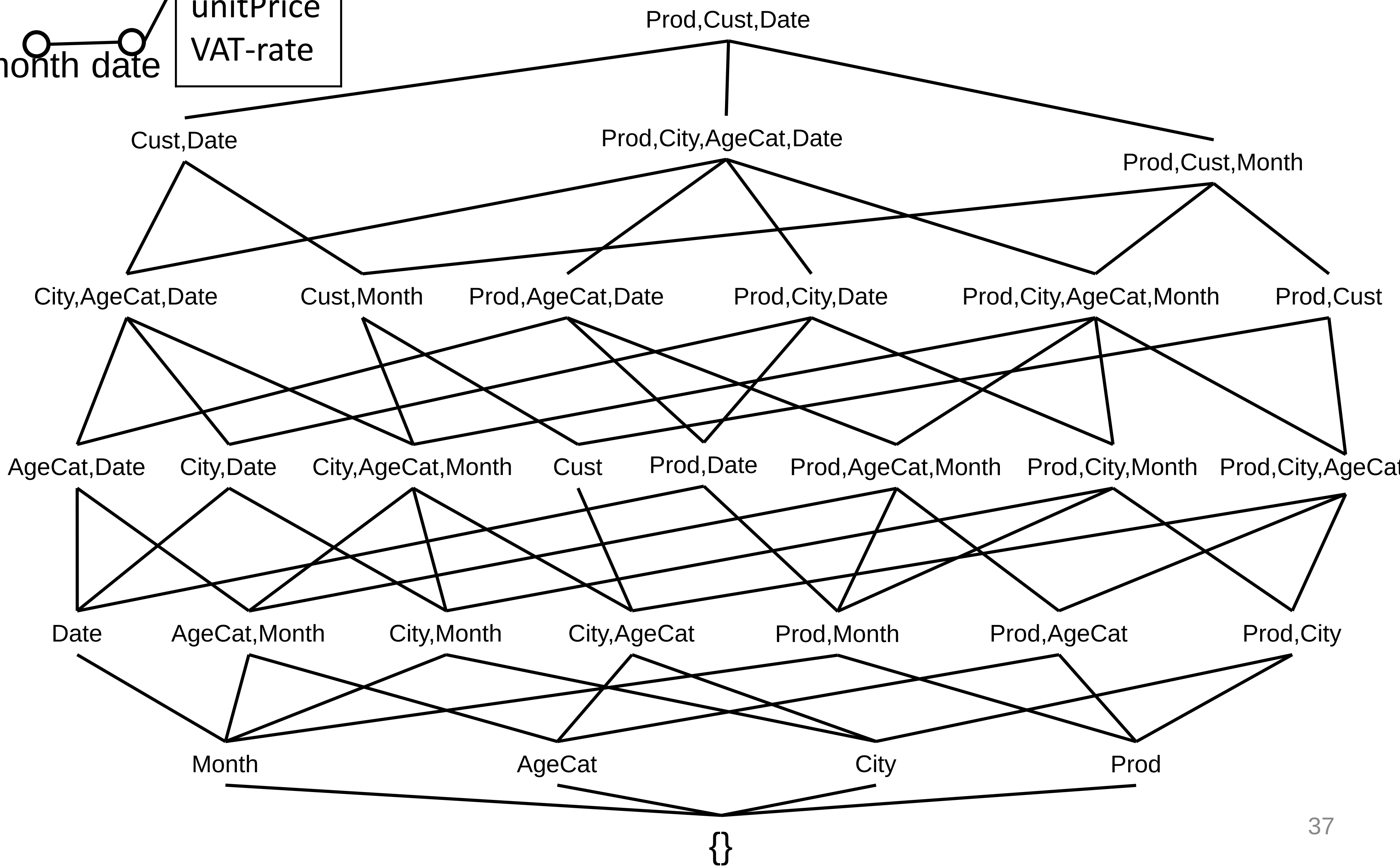


Roll-Up Lattice



Roll-Up Lattice



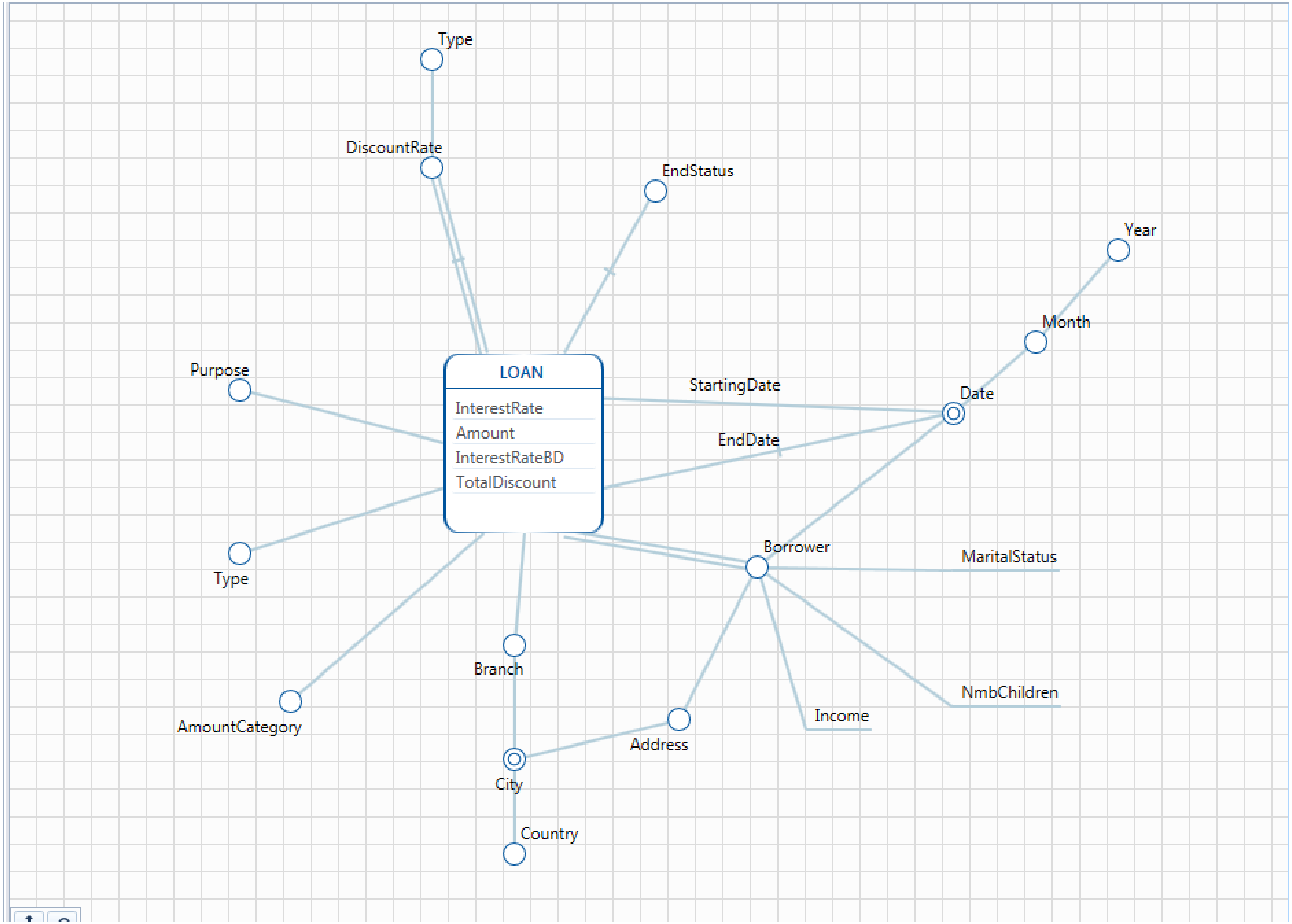


Loan Data Warehouse

A bank wants to build a data warehouse for storing and analyzing data about all loans issued by them. Every loan has one or more borrowers, a starting date, a type (e.g., fixed rate or one of different types of variable rate), the branch of the bank where the loan was issued, the interest rate at the start of the loan, and the amount. For every loan the purpose of the loan is recorded; e.g., to buy a car, a house, a personal loan, ... When a borrower applies for the loan, different discounts on the interest rate may be awarded; e.g., fidelity discount, discount because the borrower also bought some additional insurances, VIP discount, etc. For one loan, multiple discounts may apply. The amount of discount is independent of the branch. Every discount that has been awarded needs to be stored. When the loan ends, this is stored as well, together with an indication if the loan was fully repaid or the borrower defaulted. For the borrowers, their date of birth, family status, monthly income, number of children and address is stored. Throughout the lifetime of the loan the borrower make payments. Frequency and amount of the payments can vary depending on the type of loan. These payments have to be recorded as well. Sometimes a borrower may be unable to make a required payment in time. Such payment delay has to be recorded as well.

The following questions are prototypical for the type of query analysts want to answer based on the data warehouse:

- Give average interest rate before discount at the start of the loan, per loan type and branch.
- For all branches, give minimum, maximum and average interest rate per loan type and purpose.
- Give the number of loans per branch and per amount category. The amount category depends on predefined thresholds; amounts are divided into the following classes: *very high*, *high*, *medium*, *low*, and *very low*.
- Give the percentage of defaulted loans per year and per city of the branch where the loan was issued.

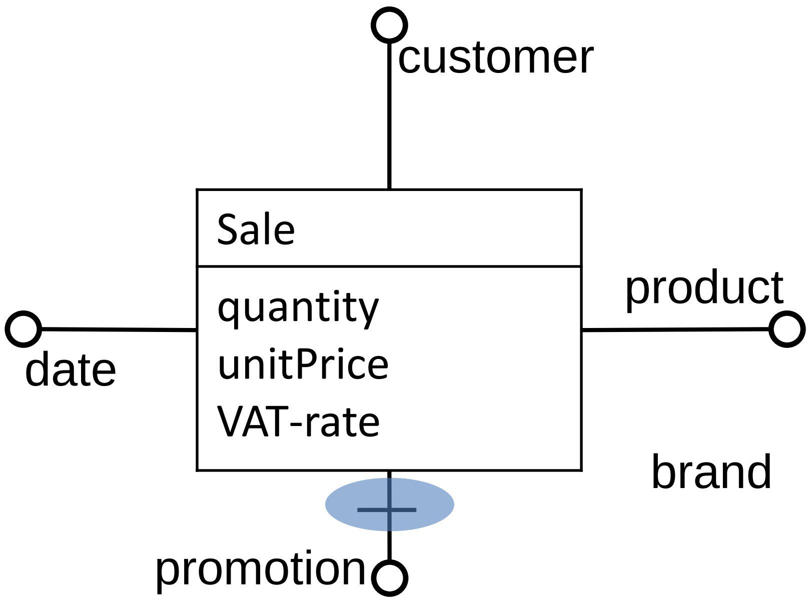

Type
DiscountRate
EndStatus
Year
Month
Purpose
LOAN
InterestRate
Amount
InterestRateBD
TotalDiscount
StartingDate
Date
EndDate
Type
Borrower
MaritalStatus
Branch
AmountCategory
NmbChildren
Income
Address
City
Country

Outline

- Special aggregation cases
- Additive and non-additive measures
- Logical Database Design
  - Star schema
  - Snowflake schema

Chapter 5 & 8 of Golfarelli & Rizzi

Optional Dimension

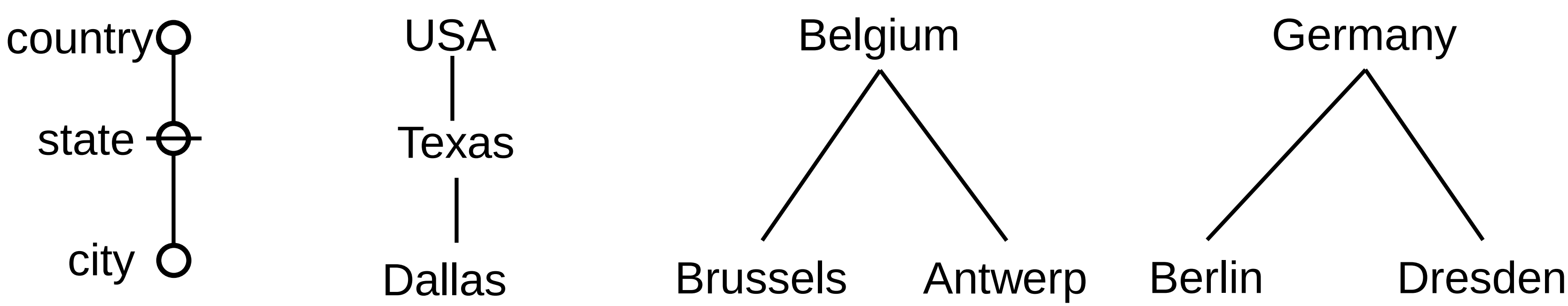


Month	Promotion	QTY
January	Discount 1	10
	Discount 2	5
	-	13
February	Discount 1	2
	-	7

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

Incomplete Hierarchies

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



Incomplete Hierarchies

Country	State	City	QTY
USA	Texas	Dallas	1
Belgium	-	Brussels	5
	-	Antwerp	5
Germany	-	Berlin	3
	-	Dresden	8

⇩ Roll-up to State

?

Incomplete Hierarchies: Solution 1

Country	State	City	QTY
USA	Texas	Dallas	1
Belgium	-	Brussels	5
	-	Antwerp	5
Germany	-	Berlin	3
	-	Dresden	8

Roll-up to State

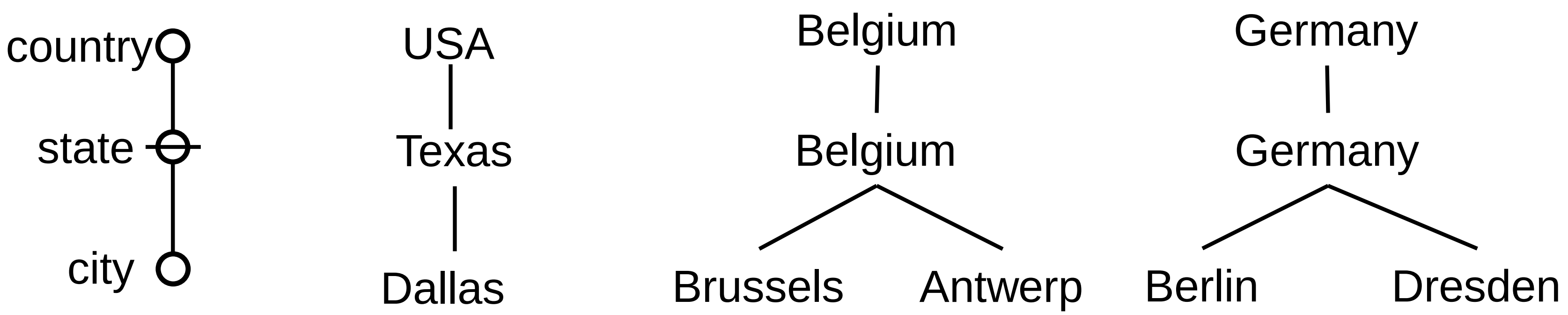
State	QTY
Texas	1
Other	21

OR

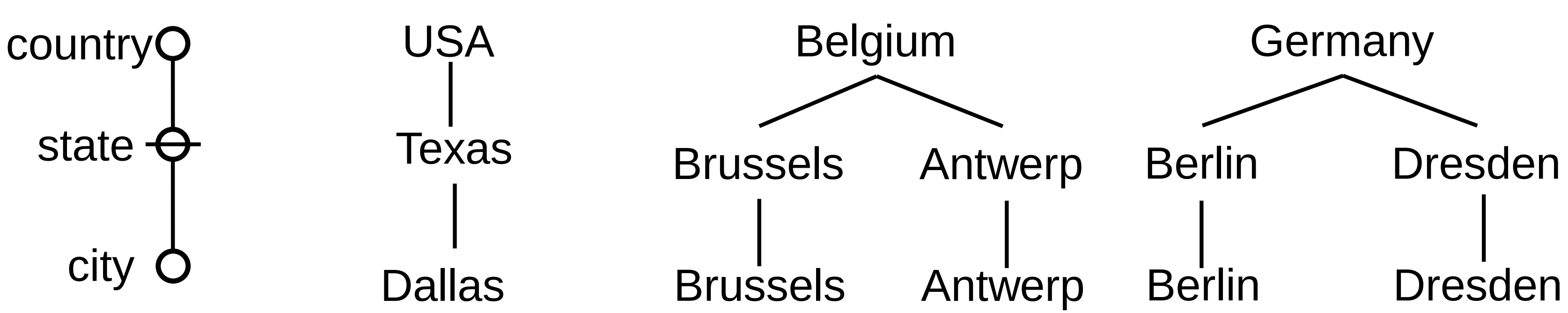
Country	State	QTY
USA	Texas	1
Belgium	Other	10
Germany	Other	11

Balancing

Upward balancing



Downward balancing



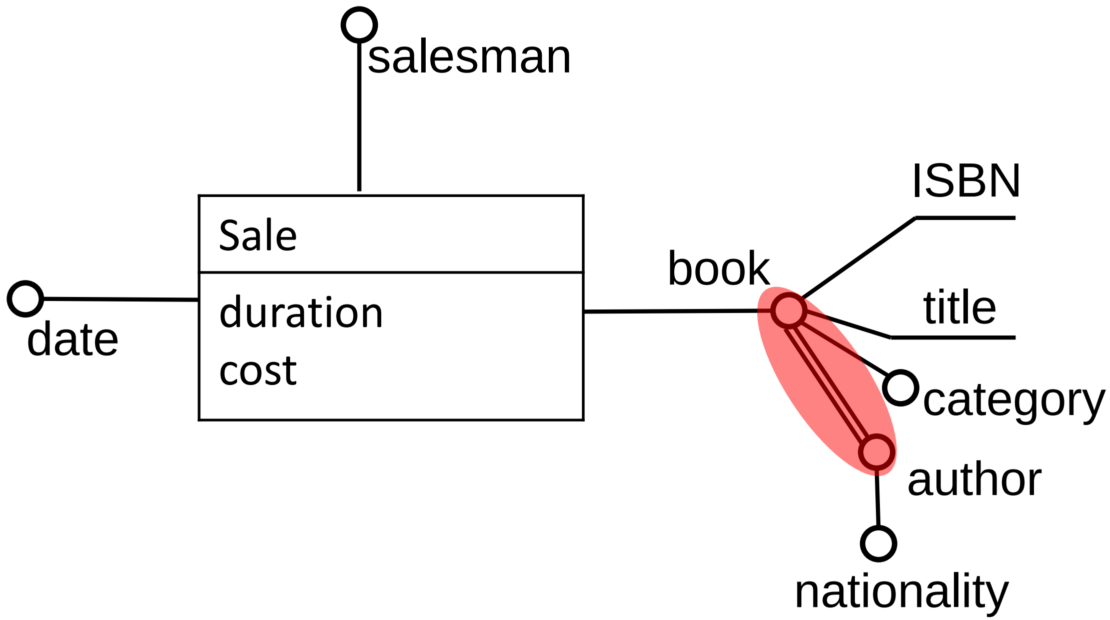
Upward Balancing

Country	State	City	QTY
USA	Texas	Dallas	1
Belgium	Belgium	Brussels	5
		Antwerp	5
Germany	Germany	Berlin	3
		Dresden	8

Roll-up to State

Country	State	QTY
USA	Texas	1
Belgium	Belgium	10
Germany	Germany	11

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

Book	Author
B1	A1
	A2
B2	A1
	A3

Author	Count()
A1	2
A2	1
A3	1

Total: 4

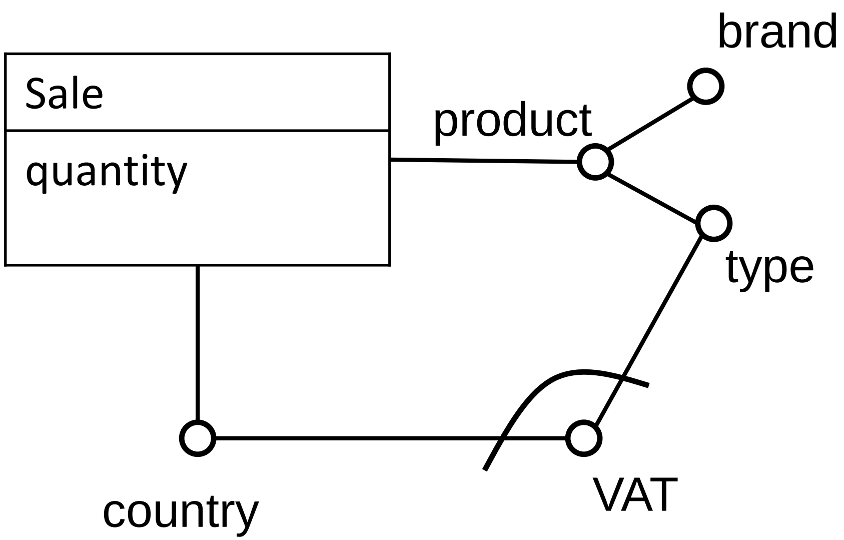
Multiple Arcs: with weights

Book	Author	Weight
B1	A1	0.5
	A2	0.5
B2	A1	0.7
	A3	0.3

Author	Weighted Count()
A1	1.2
A2	0.5
A3	0.3

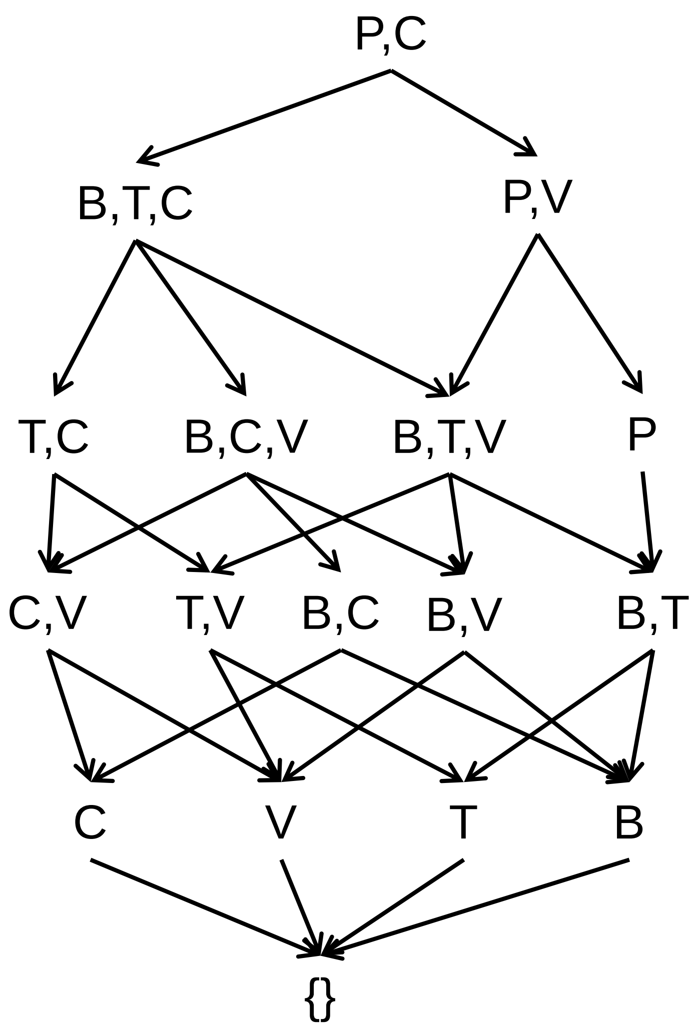
Total: 2

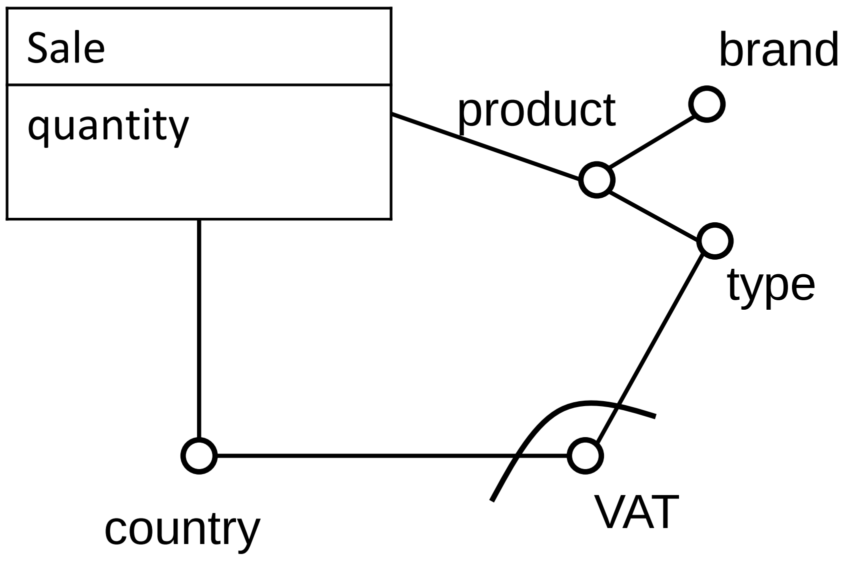
Cross-Dimensional Attributes



- VAT results in extra levels in roll-up
  - VAT; VAT,country; VAT,type
  - ~~VAT,country,type~~ or ~~VAT,country,product~~ (redundant)

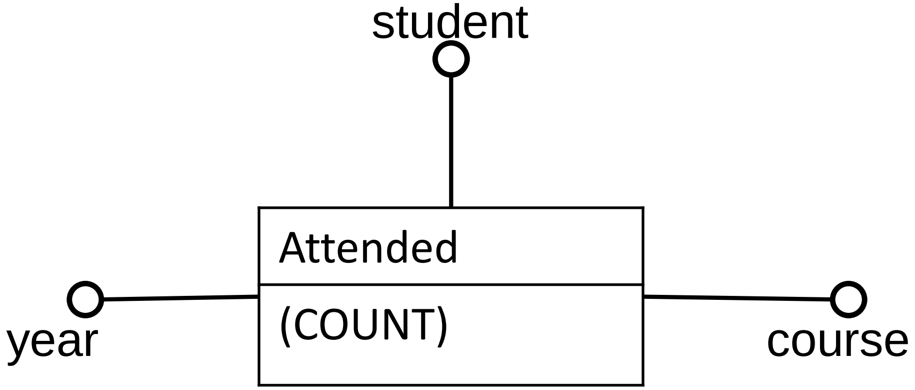
Cross-Dimensional Attributes





Measureless Schemas

- Some schemas do not have measures, but it could still be interesting to aggregate
  - COUNT; AND; OR



Measureless Schemas

		BPM	DW	ADB	DBSA
2012	John	X	X	X	X
	Mary				
	Pete				X
	Patrick	X	X		
2013	Patrick			X	X
	Jane		X		X
	Pete			X	

Measureless Schemas

COUNT

	BPM	DW	ADB	DBSA
2012	2	2	1	2
2013	0	1	2	2

OR

	BPM	DW	ADB	DBSA
John	X	X	X	X
Mary				
Pete			X	X
Patrick	X	X	X	X
Jane		X		X

Measureless Schemas

Combination of OR on Year and AND on Course

	AND
John	1
Mary	0
Pete	0
Patrick	1
Jane	0

Outline

- Special aggregation cases
- Additive and non-additive measures
- Logical Database Design
  - Star schema
  - Snowflake schema

Chapter 5 & 8 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

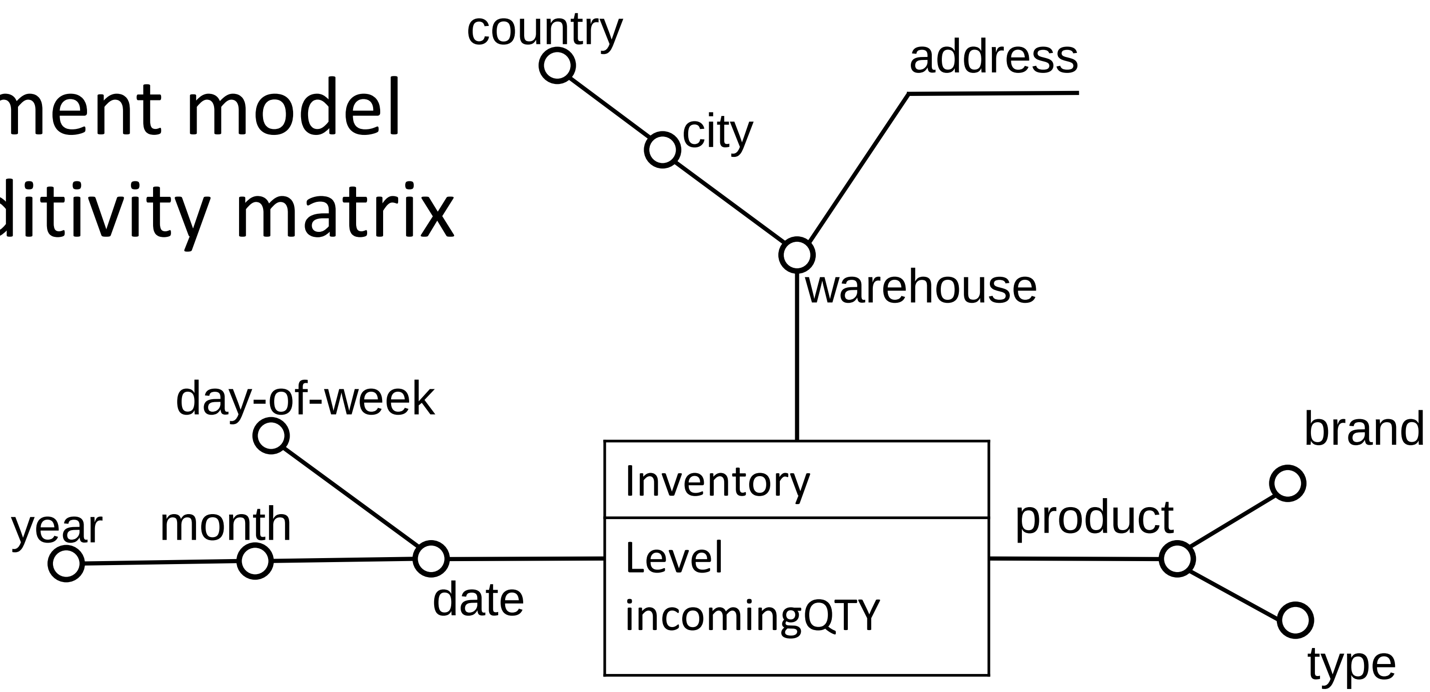
Example of non-additive measures:

Stock level over time

Unit price over time or customer

Non-Additive Measures

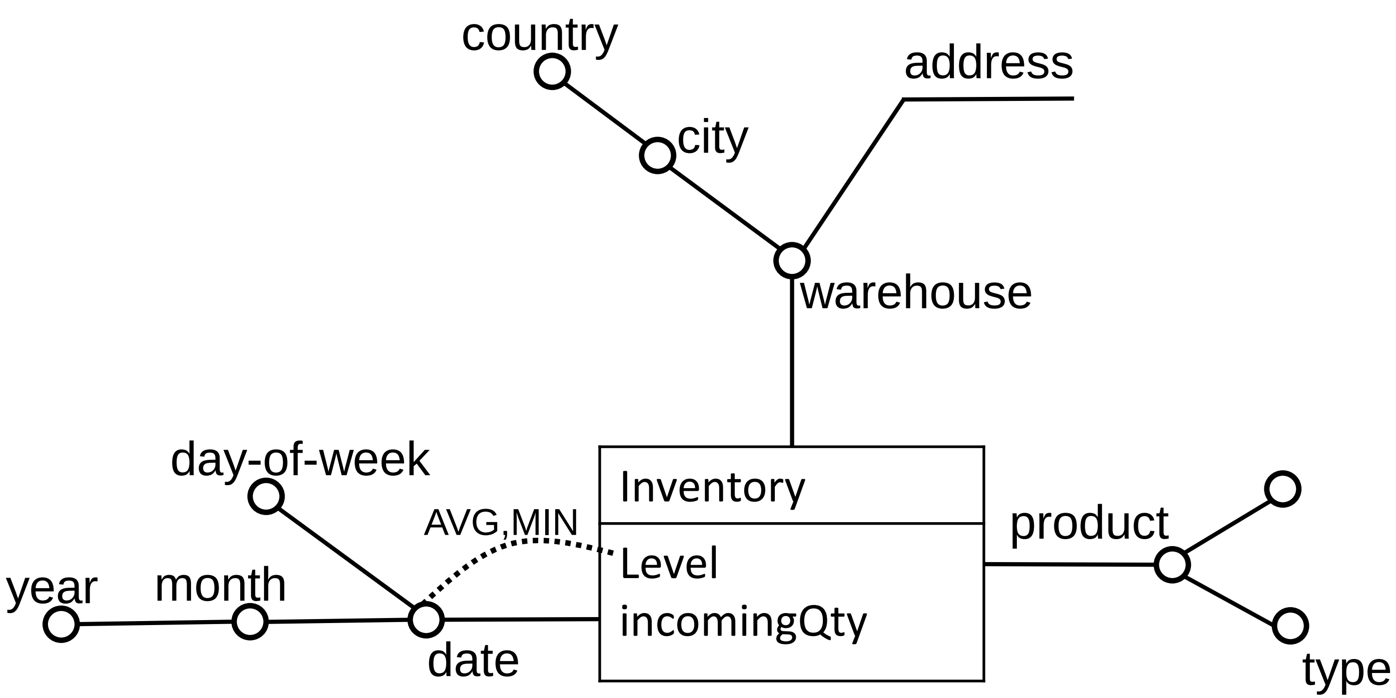
- Complement model with additivity matrix



	date	warehouse	product
Level	AVG, MIN, MAX	SUM, AVG, MIN, MAX	SUM, AVG, MIN, MAX
IncomingQTY	SUM, AVG, MIN, MAX	SUM, AVG, MIN, MAX	SUM, AVG, MIN, MAX

Non-Additive Measures

- Or, indicate directly in the schema



Distributive Operators

- SUM, MIN, MAX are distributive aggregation operators
  - SUM(A,B,C,D) = SUM(SUM(A,B),Sum(C,D)) = SUM(SUM(A,B,C),D) = …

- For distributive operators it holds:
  - If G1 above G2 in roll-up lattice, then group by G2 can be computed directly from group by G1

Example: Distributive Operator

Continent	Country	City	Amount
Europe	Belgium	Brussels	5
		Antwerp	3
	Germany	Berlin	2
North-America	USA	Chicago	1
		Tampa	8

Continent	Country	SUM(Amount)
Europe	Belgium	8
	Germany	2
North-America	USA	9

Continent	Sum(Amount)
Europe	10
North-America	9

Sum(Amount)
19

Example: Non-Distributive Operator

Continent	Country	City	Amount
Europe	Belgium	Brussels	5
		Antwerp	3
	Germany	Berlin	2
North-America	USA	Chicago	1
		Tampa	8

AVG(Amount)
3.8

Continent	Country	AVG(Amount)
Europe	Belgium	4
	Germany	2
North-America	USA	4.5

Continent	AVG(Amount)
Europe	3.33
North-America	4.5

Algebraic Operator

- With some additional information some non-distributive aggregation operators can still be calculated from partial aggregates
  - AVG(A,B,C,D) = SUM(A,B,C,D) / COUNT(A,B,C,D)
  - VAR(A,B,C,D) = AVG(A^2,B^2,C^2,D^2) - 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