Data representation

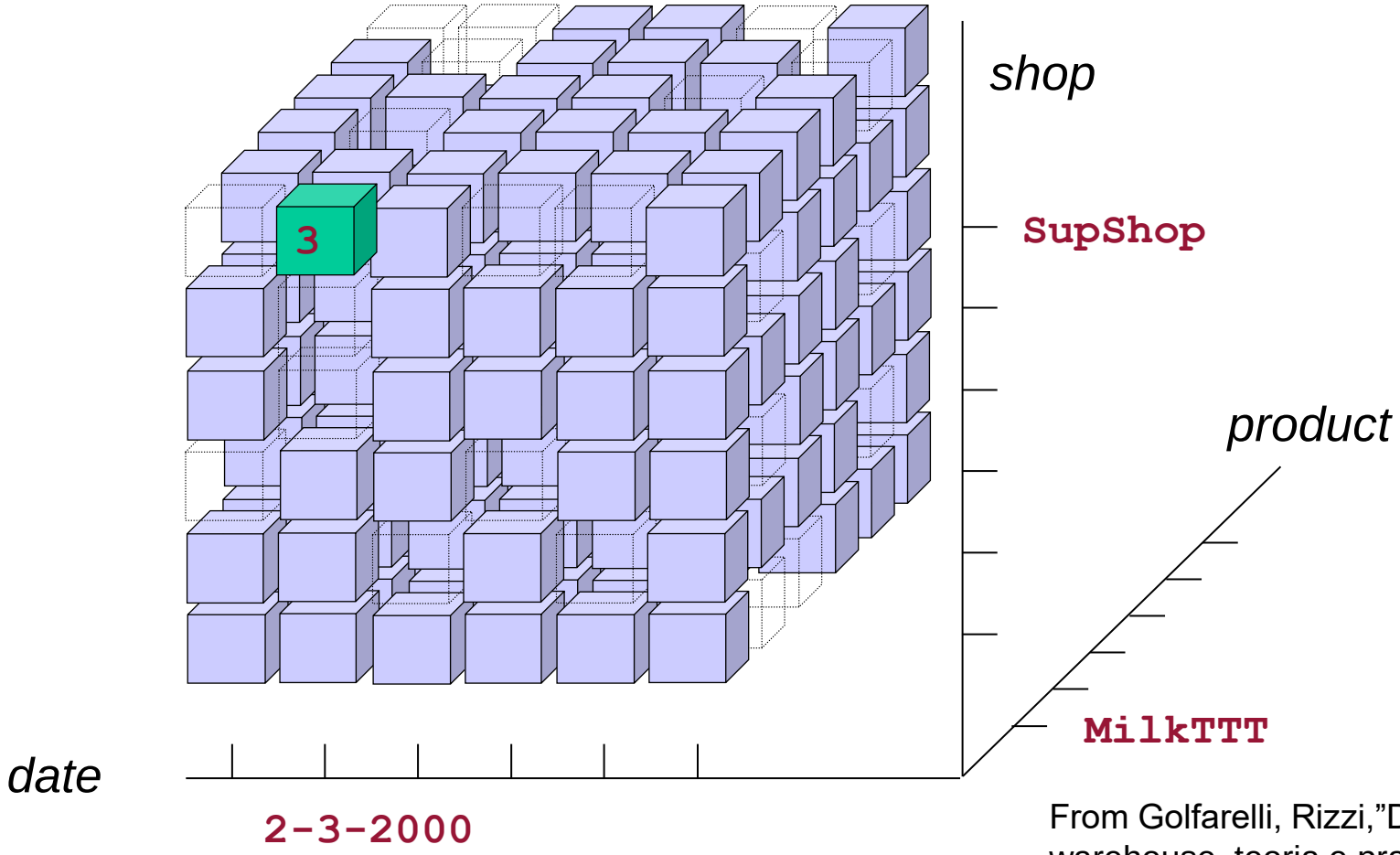
Elena Baralis
Politecnico di Torino

Multidimensional data representation

- Data are represented as an (hyper)cube with three or more dimensions
- Measures on which analysis is performed: cells at dimension intersection
- Data warehouse for tracking sales in a supermarket chain:
  - dimensions: product, shop, time
  - measures: sold quantity, sold amount, ...

Elena Baralis
Politecnico di Torino

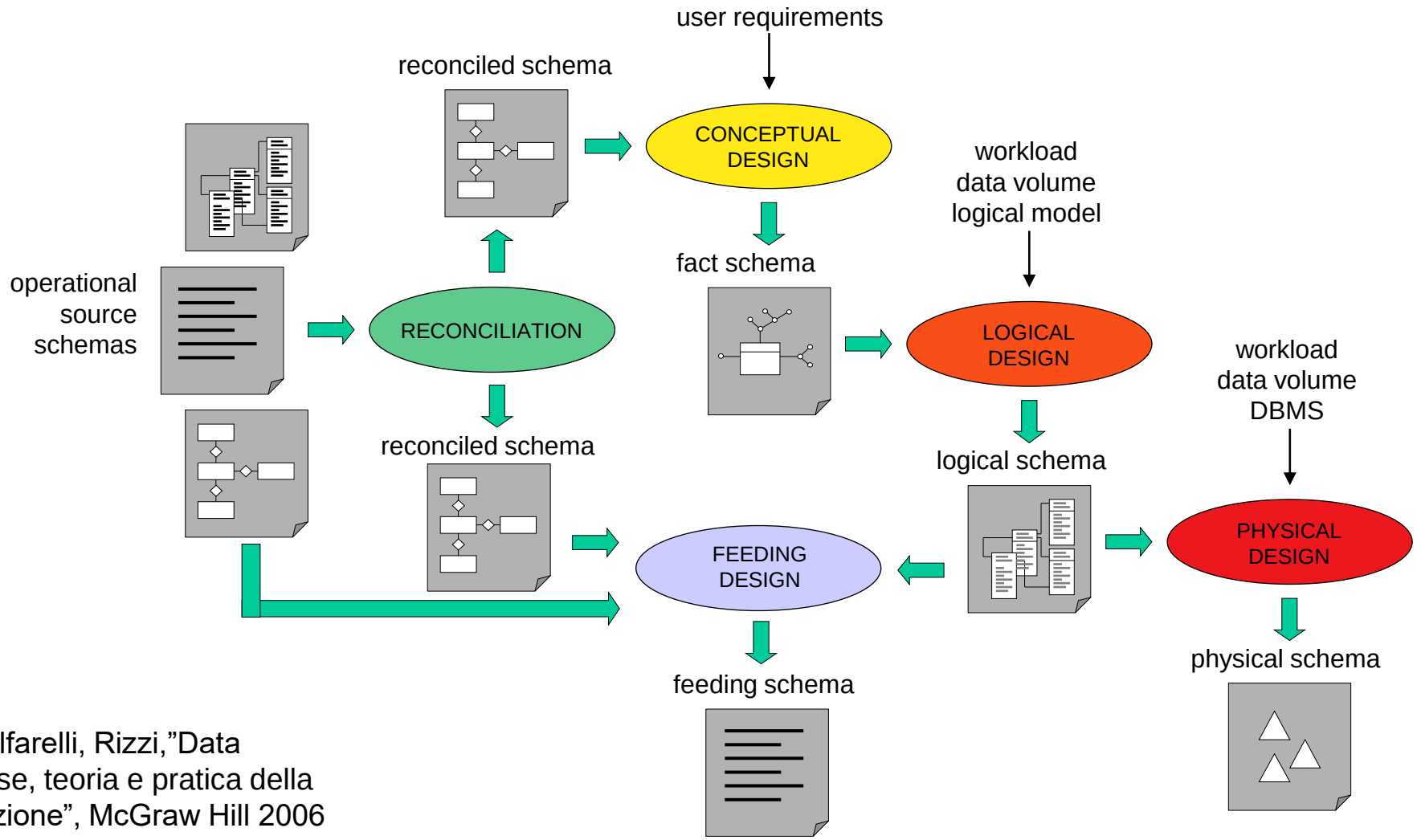
Multidimensional representation



From Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Data mart design



From Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Conceptual design

Elena Baralis
Politecnico di Torino

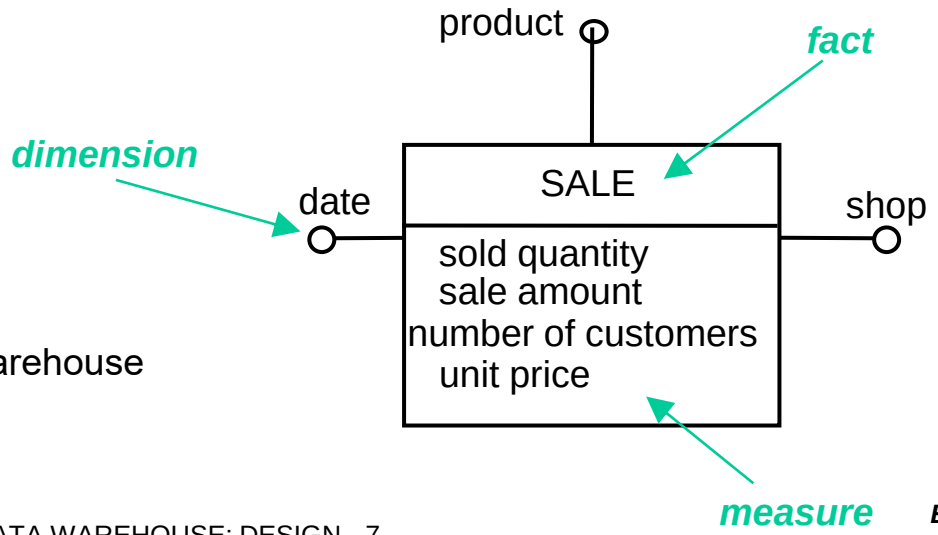
Conceptual design

- No currently adopted modeling formalism
  - ER model not adequate
- *Dimensional Fact Model* (Golfarelli, Rizzi)
  - graphical model supporting conceptual design
  - for a given fact, it defines a *fact schema* modelling
    - dimensions
    - hierarchies
    - measures
  - it provides design documentation both for requirement review with users, and after deployment

Elena Baralis
Politecnico di Torino

Dimensional Fact Model

- Fact
  - it models a set of relevant events (sales, shippings, complaints)
  - it evolves with time
- Dimension
  - it describes the *analysis coordinates of a fact* (e.g., each sale is described by the sale date, the shop and the sold product)
  - it is characterized by many, typically categorical, attributes
- Measure
  - it describes a *numerical property of a fact* (e.g., each sale is characterized by a sold quantity)
  - aggregates are frequently performed on measures

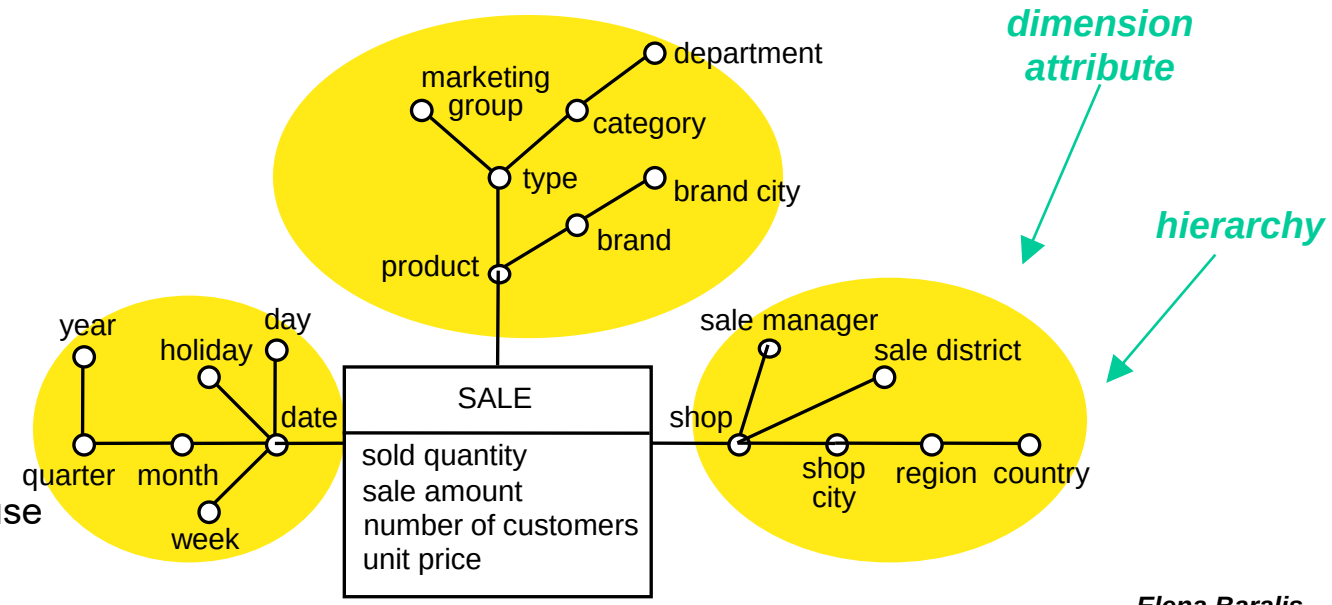


From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

Elena Baralis
Politecnico di Torino

DFM: Hierarchy

- Each dimension can have a set of associated attributes
- The attributes describe the dimension at *different abstraction levels* and can be structured as a *hierarchy*
- The hierarchy represents a *generalization relationship* among a subset of attributes in a dimension (e.g., geografic hierarchy for the shop dimension)
- The hierarchy represents a functional dependency (1:n relationship)

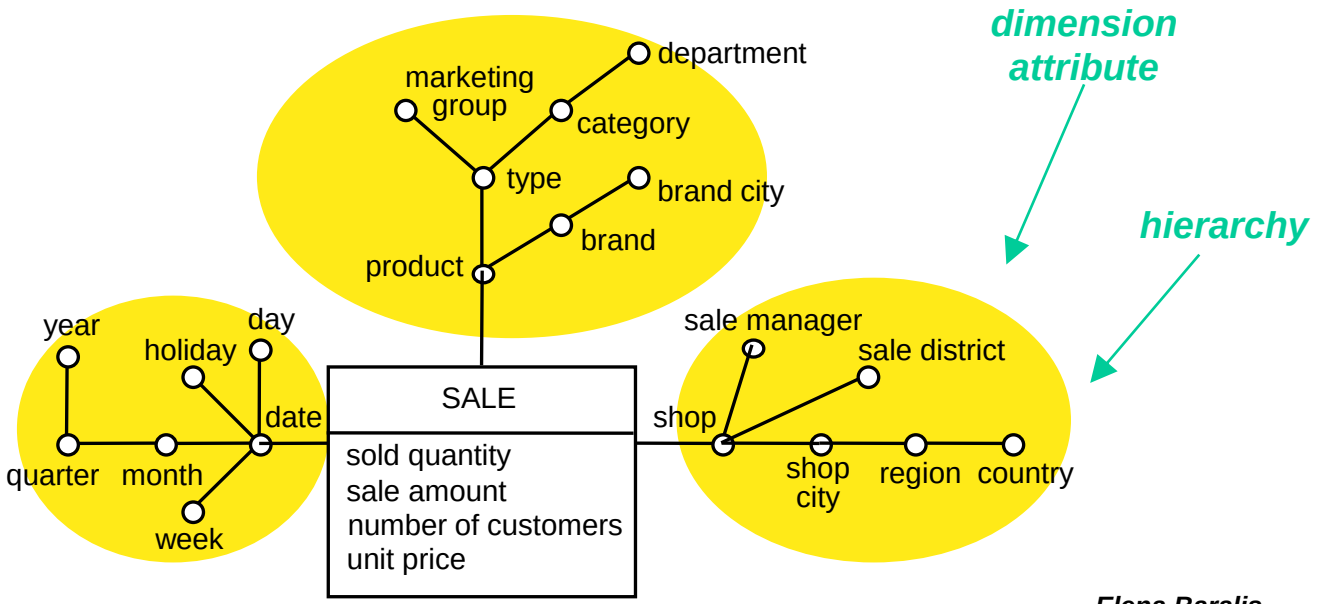


From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

Elena Baralis
Politecnico di Torino

Aggregation

- Aggregation computes measures with a coarser granularity than those in the original fact schema
  - detail reduction is usually obtained by *climbing a hierarchy*
  - standard aggregate operators: SUM, MIN, MAX, AVG, COUNT



Elena Baralis
Politecnico di Torino

Example aggregation patterns

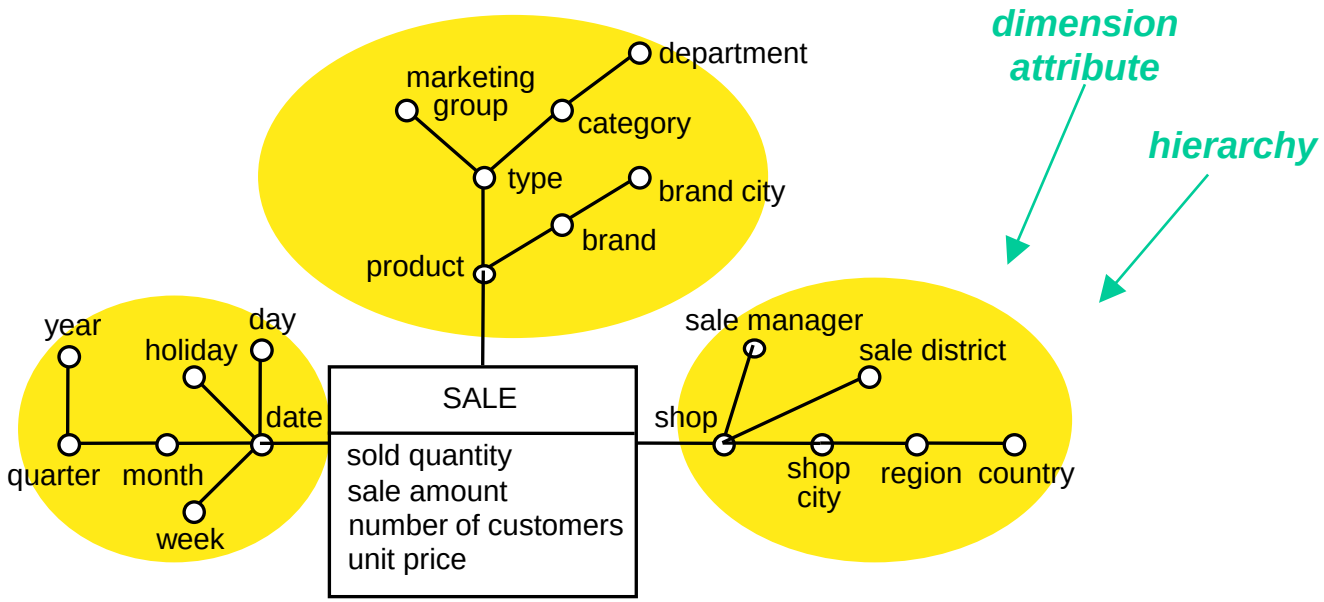
(date, product, shop)

(date, product, shop city)

(month, product, shop city)

(month, type, shop city)

(year, type, country)



Elena Baralis
Politecnico di Torino

Example aggregation patterns

category	type	product
home cleaning	washing powder	Brillo
		Sbianco
		Lucido
	soap	Manipulite
		Scent
food	milk	Latte F Slurp
		Latte U Slurp
		Yogurt Slurp
	soda	Bevimi
		Colissima

year / quart.	1999 I '99	1999 II '99	1999 III '99	1999 IV '99	2000 I '00	2000 II '00	2000 III '00	2000 IV '00
	100	90	95	90	80	70	90	85
	20	30	20	10	25	30	35	20
	60	50	60	45	40	40	50	40
	15	20	25	30	15	15	20	10
	30	35	20	25	30	30	20	15
	90	90	85	75	60	80	85	60
	60	80	85	60	70	70	75	65
	20	30	40	35	30	35	35	20
	20	10	25	30	35	30	20	10
	50	60	45	40	50	60	45	40

Measure: sold quantity

Aggregate operator: SUM

year / quart.	1999 I'99	1999 II'99	1999 III'99	1999 IV'99	2000 I'00	2000 II'00	2000 III'00	2000 IV'00
category								
home clean.	225	225	220	200	190	185	215	170
food	240	270	280	240	245	275	260	195

year	1999	2000
category		
home clean.	870	760
food	1030	975

category	type	year 1999	year 2000
home cleaning	washing p.	670	605
	soap	200	155
food	milk	750	685
	soda	280	290

From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

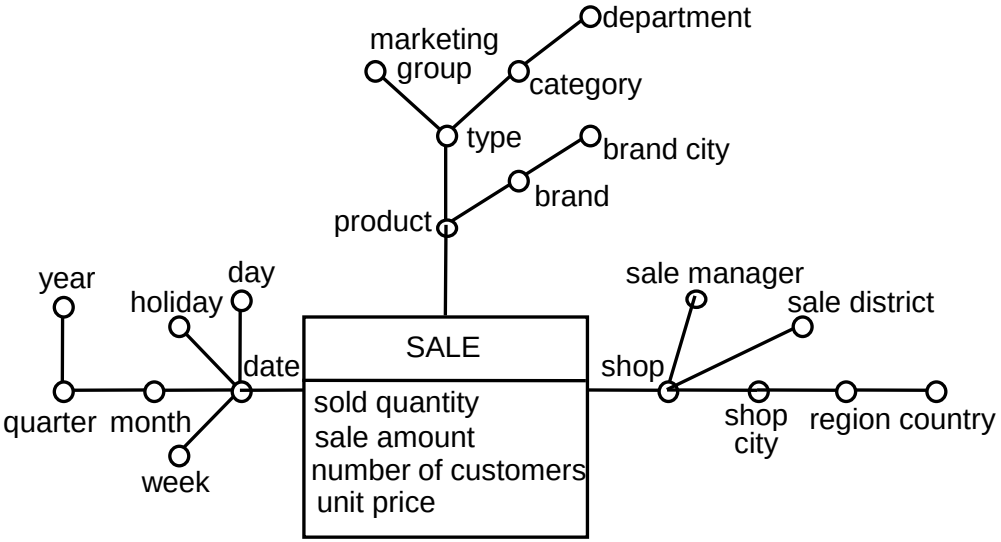
Elena Baralis
Politecnico di Torino

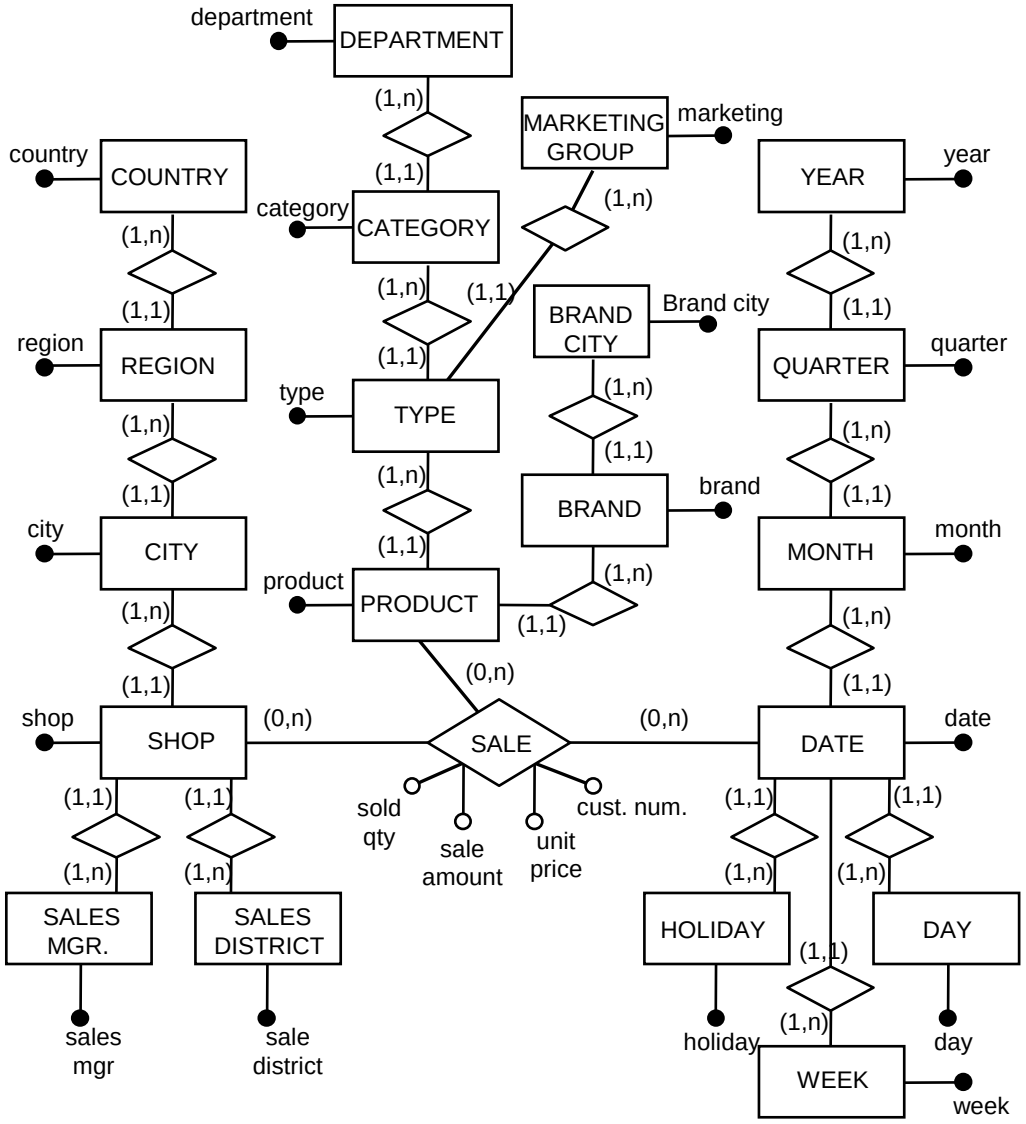
Measure classification

- Stream measures
  - can be evaluated cumulatively at the end of a time period
  - can be aggregated by means of all standard operators
  - examples: sold quantity, sale amount
- Level measures
  - evaluated at a given time (snapshot)
  - not additive along the time dimension
  - examples: inventory level, account balance
- Unit measures
  - evaluated at a given time and expressed in relative terms
  - not additive along any dimension
  - examples: unit price of a product

Elena Baralis
Politecnico di Torino

Comparison between DFM and ER





From Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Logical design

Elena Baralis
Politecnico di Torino

Logical design

- We address the relational model (ROLAP)
  - inputs
    - conceptual fact schema
    - workload
    - data volume
    - system constraints
  - output
    - relational logical schema
- Based on different principles with respect to traditional logical design
  - data redundancy
  - table denormalization

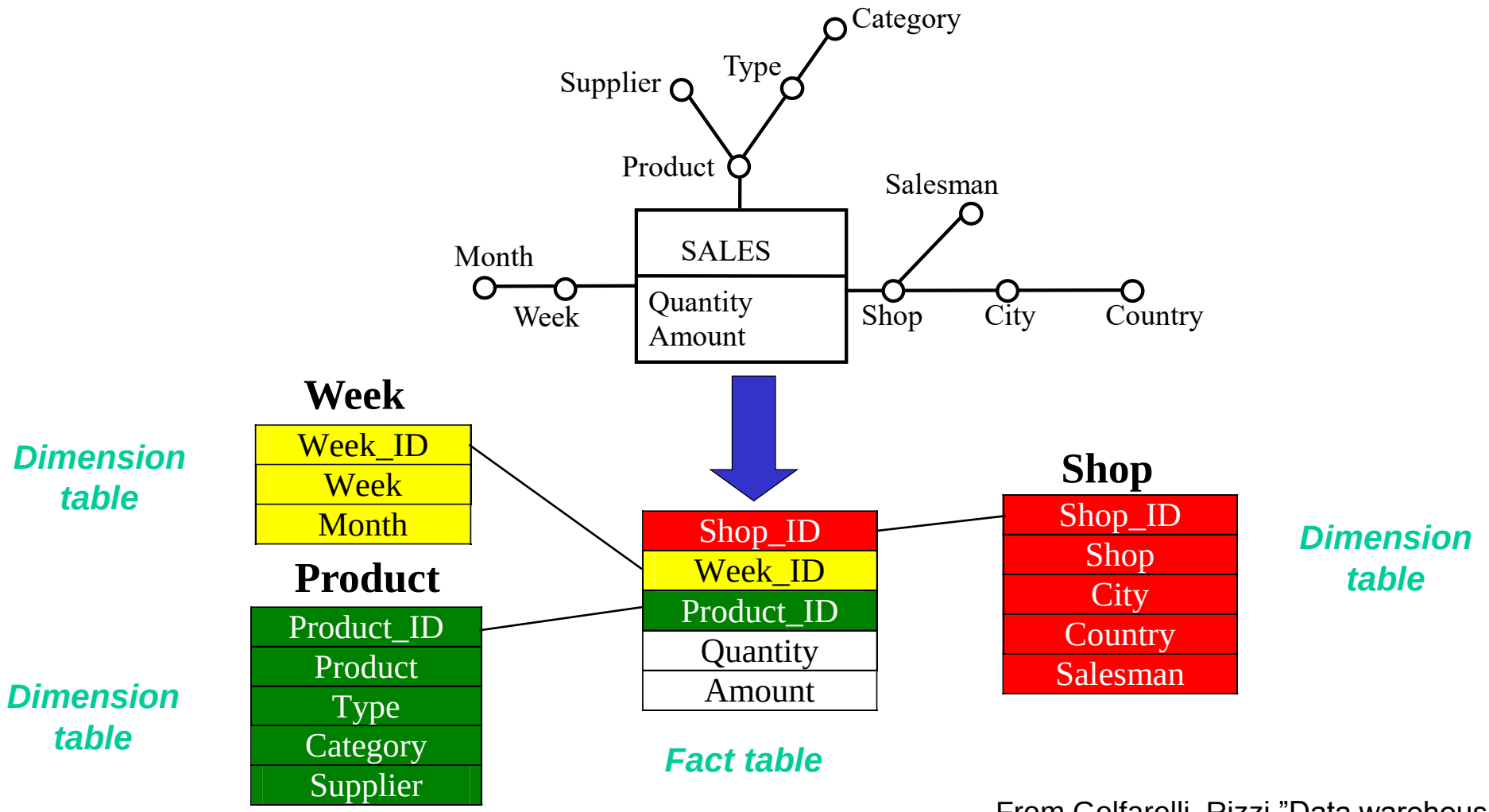
Elena Baralis
Politecnico di Torino

Star schema

- Dimensions
  - one table for each dimension
  - surrogate (generated) primary key
  - it contains all dimension attributes
  - hierarchies are not explicitly represented
    - all attributes in a table are at the same level
  - totally denormalized representation
    - it causes data redundancy
- Facts
  - one fact table for each fact schema
  - primary key composed by foreign keys of all dimensions
  - measures are attributes of the fact table

Elena Baralis
Politecnico di Torino

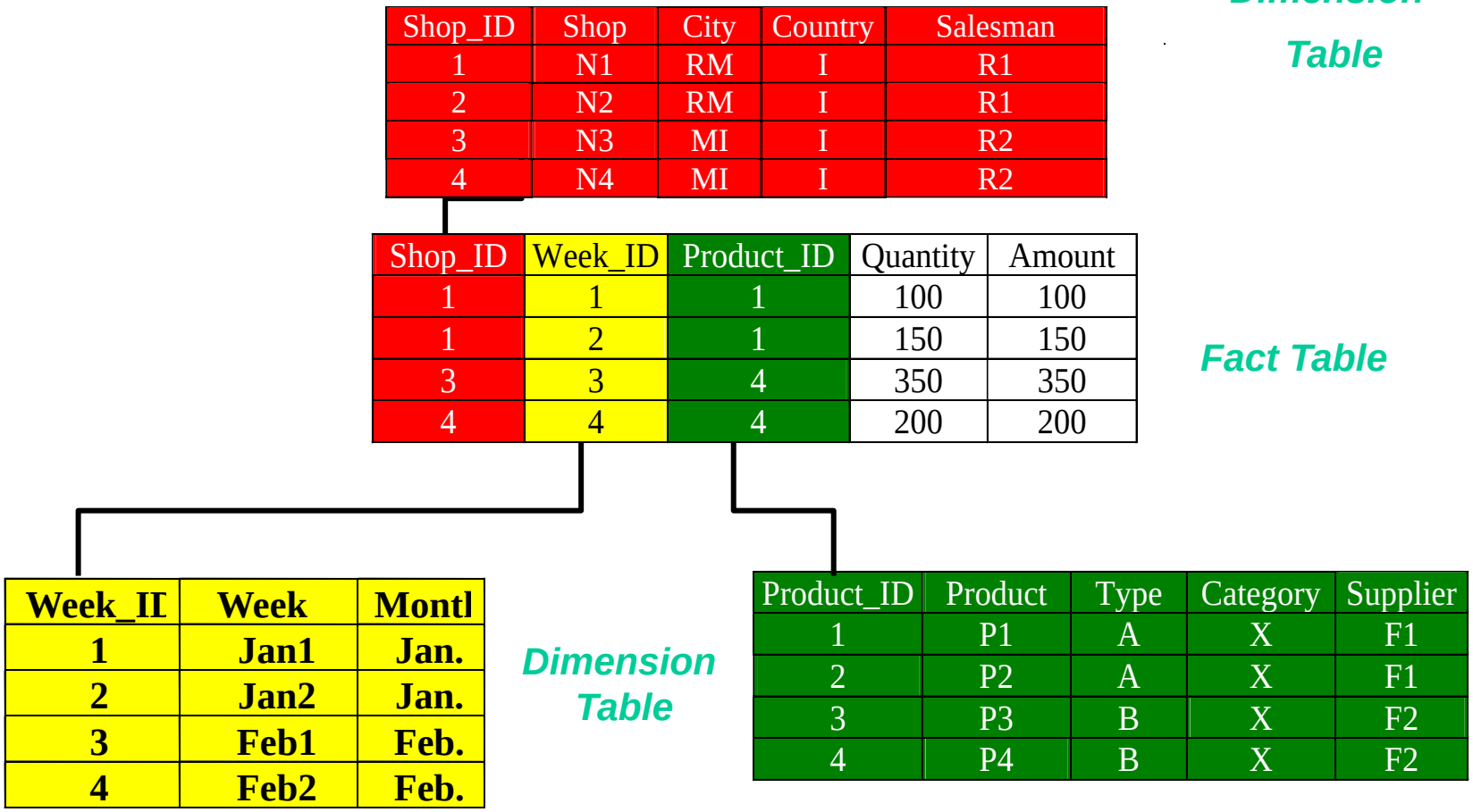
Star schema



From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

Elena Baralis
Politecnico di Torino

Star schema



Shop_ID	Shop	City	Country	Salesman
1	N1	RM	I	R1
2	N2	RM	I	R1
3	N3	MI	I	R2
4	N4	MI	I	R2

Shop_ID	Week_ID	Product_ID	Quantity	Amount
1	1	1	100	100
1	2	1	150	150
3	3	4	350	350
4	4	4	200	200

Week_ID	Week	Montl
1	Jan1	Jan.
2	Jan2	Jan.
3	Feb1	Feb.
4	Feb2	Feb.

Product_ID	Product	Type	Category	Supplier
1	P1	A	X	F1
2	P2	A	X	F1
3	P3	B	X	F2
4	P4	B	X	F2

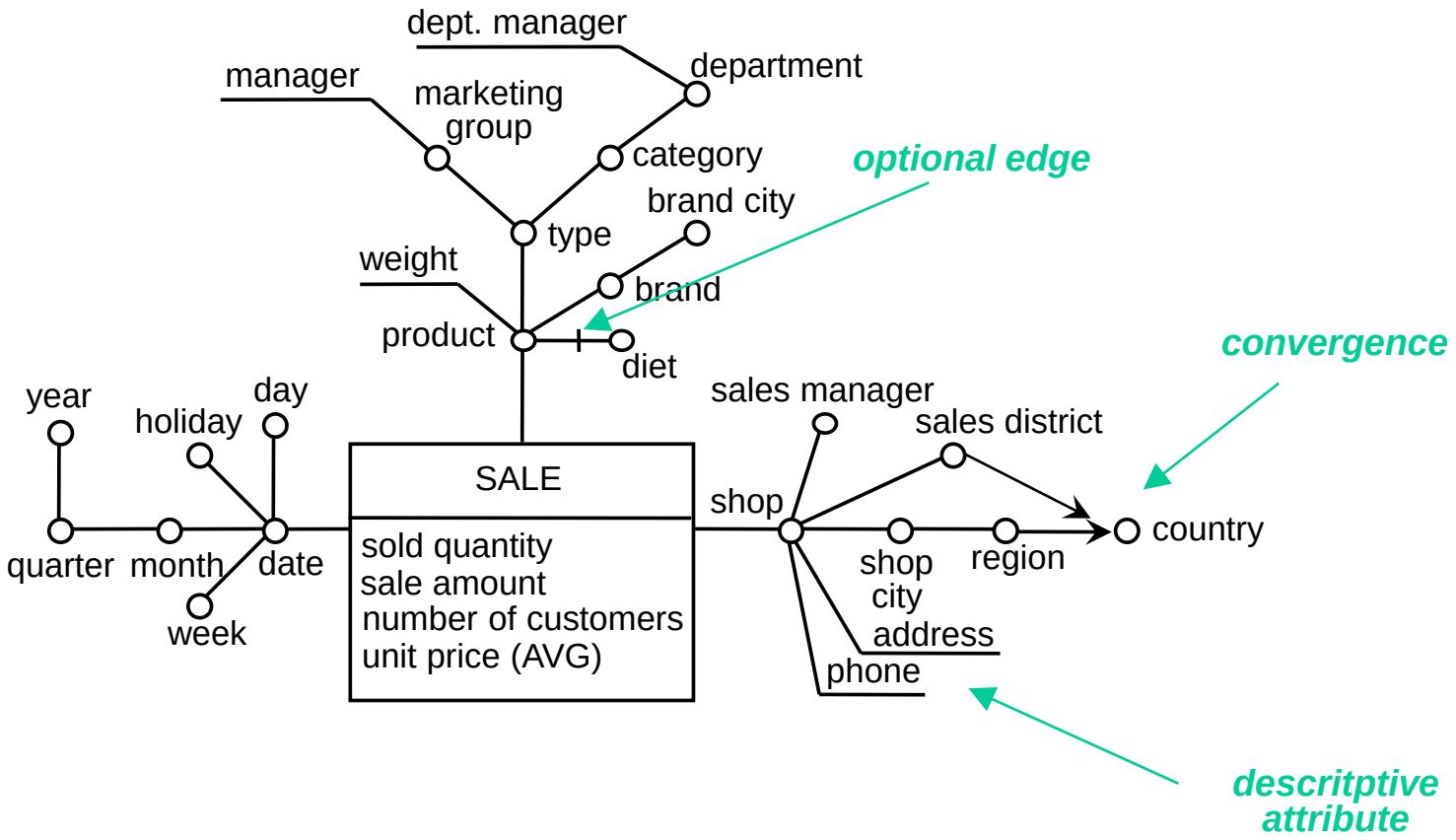
From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

Elena Baralis
Politecnico di Torino

Advanced DFM

Elena Baralis
Politecnico di Torino

Advanced DFM

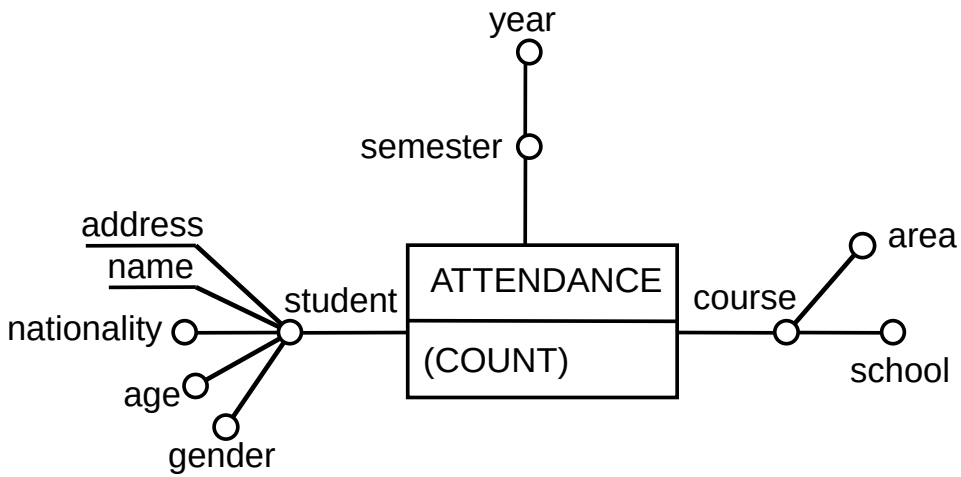


From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

Elena Baralis
Politecnico di Torino

Factless fact schema

- Some events are not characterized by measures
  - empty (i.e., factless) fact schema
  - it records occurrence of an event
- Used for
  - counting occurred events (e.g., course attendance)
  - representing events not occurred (coverage set)

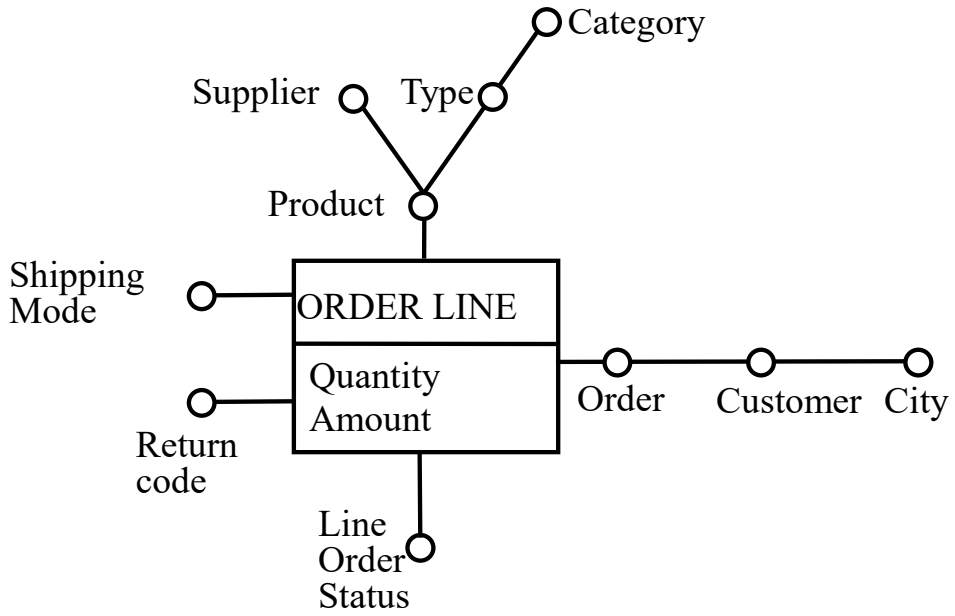


From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

Elena Baralis
Politecnico di Torino

Degenerate dimensions

- Dimensions with a single attribute



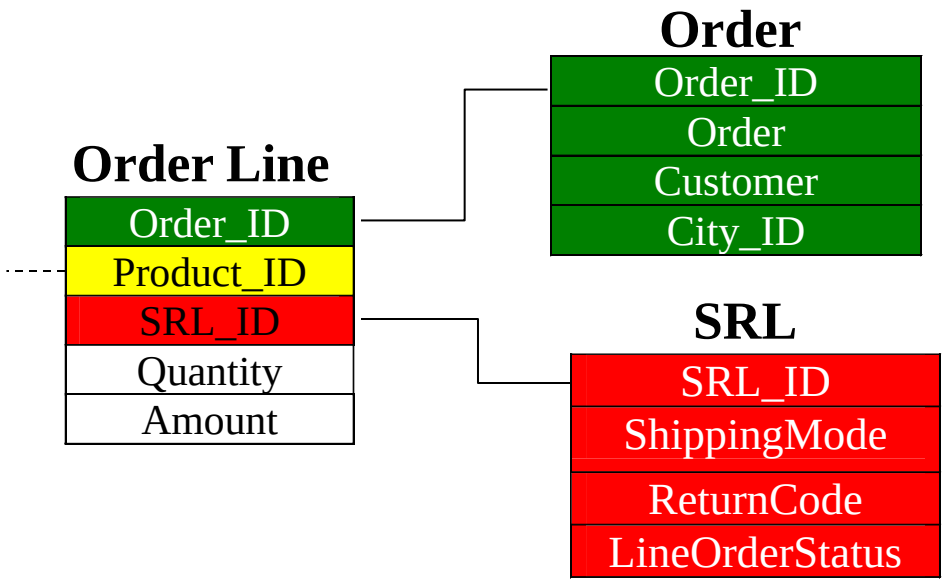
Elena Baralis
Politecnico di Torino

Degenerate dimensions

- Implementations
  - Integration into the fact table
    - for attributes with a (very) small size
  - junk dimension
    - single dimension containing several degenerate dimensions
    - no functional dependencies among attributes in the junk dimension
      - all attribute value combinations are allowed
      - feasible only for attribute domains with small cardinality

Elena Baralis
Politecnico di Torino

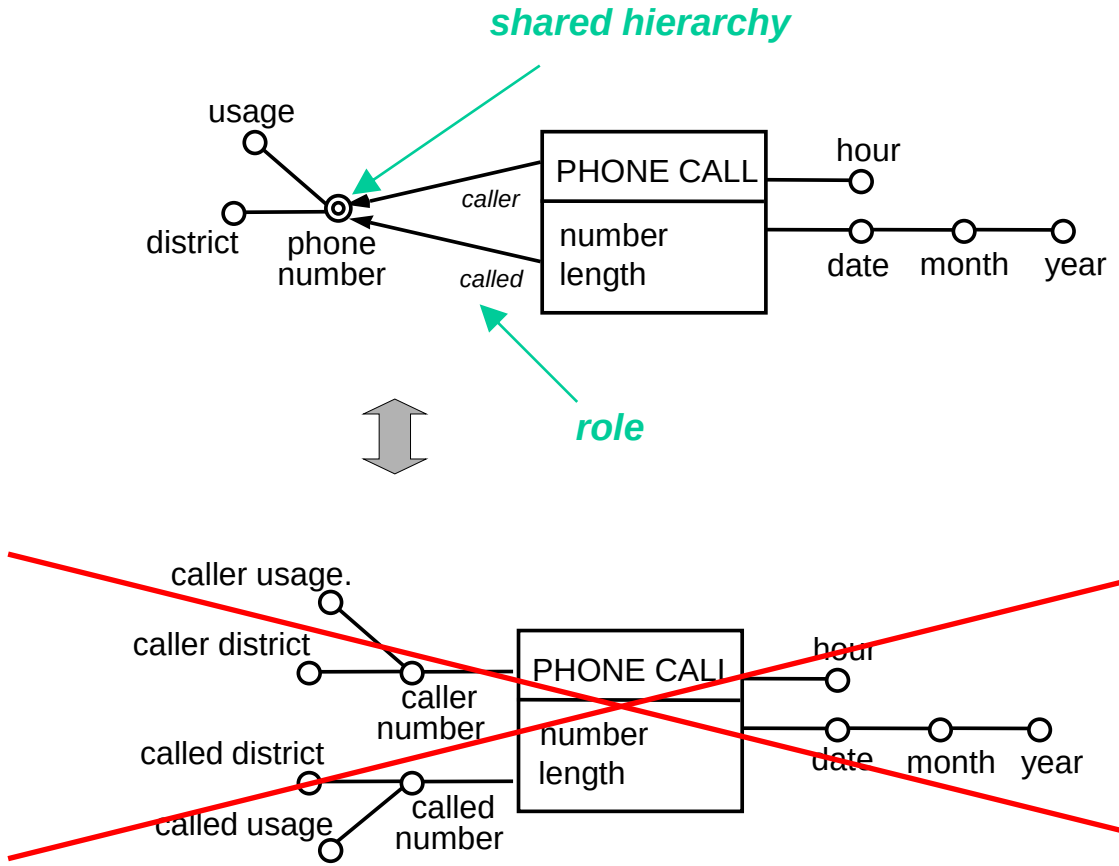
Junk dimension



From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

Elena Baralis
Politecnico di Torino

Advanced DFM



From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

Elena Baralis
Politecnico di Torino

Representing time

Elena Baralis
Politecnico di Torino

Representing time

- Data modification over time is explicitly represented by event occurrences
  - time dimension
  - events stored as facts
- Also dimensions may change over time
  - modifications are typically slower
    - slowly changing dimension [Kimball]
  - examples: client demographic data, product description
  - if required, dimension evolution should be explicitly modeled

Elena Baralis
Politecnico di Torino

How to represent time (type I)

- Snapshot of the current value
  - data is overwritten with the current value
  - it overrides the past with the current situation
  - used when an explicit representation of the data change is not needed
  - example
    - customer Mario Rossi changes marital status after marriage
    - all his purchases correspond to the “married” customer

Elena Baralis
Politecnico di Torino

How to represent time (type II)

- Events are related to the temporally corresponding dimension value
  - after each state change in a dimension
    - a new dimension instance is created
    - new events are related to the new dimension instance
  - events are partitioned after the changes in dimensional attributes
  - example
    - customer Mario Rossi changes marital status after marriage
    - his purchases are partitioned in purchases performed by “unmarried” Mario Rossi and purchases performed by “married” Mario Rossi (a new instance of Mario Rossi)

Elena Baralis
Politecnico di Torino

How to represent time (type III)

- All events are mapped to a dimension value sampled at a given time
  - it requires the explicit management of dimension changes during time
    - the dimension schema is modified by introducing
      - two timestamps: validity start and validity end
      - a new attribute which allows identifying the sequence of modifications on a given instance (e.g., a “master” attribute pointing to the root instance)
    - each state change in the dimension requires the creation of a new instance

Elena Baralis
Politecnico di Torino

How to represent time (type III)

- Example
  - customer Mario Rossi changes marital status after marriage
  - validity end timestamp of first Mario Rossi instance is given by the marriage date
  - validity start timestamp of the new instance is the same day
  - purchases are partitioned as in type II
  - a new attribute allows tracking all changes of Mario Rossi instance

Elena Baralis
Politecnico di Torino

Aggregate operators

Elena Baralis
Politecnico di Torino

Aggregate operators

- Distributive
  - can always compute higher level aggregations from more detailed data
  - examples: sum, min, max
- Non distributive operators
  - can compute higher level aggregations from more detailed data *only* when supplementary support measures are available
  - examples: avg (it requires count)

Elena Baralis
Politecnico di Torino

Distributive operators

category	type	product
home cleaning	washing powder	Brillo
		Sbianco
		Lucido
	soap	Manipulite
		Scent
food	milk	Latte F Slurp
		Latte U Slurp
		Yogurt Slurp
	soda	Bevimi
		Colissima

year	1999				2000			
quart.	I '99	II '99	III '99	IV '99	I '00	II '00	III '00	IV '00
	100	90	95	90	80	70	90	85
	20	30	20	10	25	30	35	20
	60	50	60	45	40	40	50	40
	15	20	25	30	15	15	20	10
	30	35	20	25	30	30	20	15
	90	90	85	75	60	80	85	60
	60	80	85	60	70	70	75	65
	20	30	40	35	30	35	35	20
	20	10	25	30	35	30	20	10
	50	60	45	40	50	60	45	40

year	1999				2000			
quart.	I'99	II'99	III'99	IV'99	I'00	II'00	III'00	IV'00
category								
home clean.	225	225	220	200	190	185	215	170
food	240	270	280	240	245	275	260	195

year	1999	2000
category		
home clean.	870	760
food	1030	975

year	1999	2000
category	type	
home cleaning	washing p.	670	605
	soap	200	155
food	milk	750	685
	soda	280	290

From Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Non distributive operators

category	type	product
home cleaning	washing powder	Brillo
		Sbianco
		Lucido
	soap	Manipulite
		Scent

year / quart.	1999 I'99	II'99	III'99	IV'99
	2	2	2,2	2,5
	1,5	1,5	2	2,5
	–	3	3	3
	1	1,2	1,5	1,5
	1,5	1,5	2	–

Measure: unit price

category	type
home cleaning	wash. p.
	soap
	avg:

year / quart.	1999 I'99	II'99	III'99	IV'99
	1,75	2,17	2,40	2,67
	1,25	1,35	1,75	1,50
	1,50	1,76	2,08	2,09

category	year / quart.	1999 I'99	II'99	III'99	IV'99
home clean.		1,50	1,84	2,14	2,38

From Golfarelli, Rizzi,"Data warehouse design", McGraw Hill

Elena Baralis
Politecnico di Torino