

**Data warehouse design**

Elena Baralis  
Politecnico di Torino

Database and data mining group, Politecnico di Torino  
Copyright - All rights reserved  
DATA WAREHOUSE: DESIGN - 1

## Risk factors

- High user expectation
  - the data warehouse is *the* solution of the company's problems
- Data and OLTP process quality
  - incomplete or unreliable data
  - non integrated or non optimized business processes
- "Political" management of the project
  - cooperation with "information owners"
  - system acceptance by end users
  - deployment
    - appropriate training

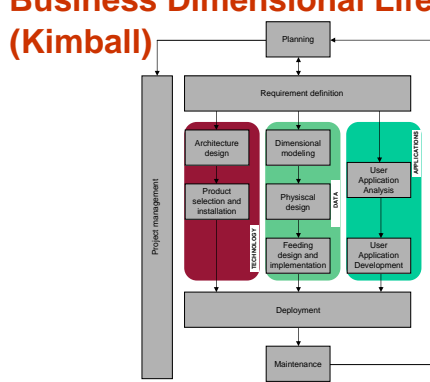
Database and data mining group, Politecnico di Torino  
Copyright - All rights reserved  
DATA WAREHOUSE: DESIGN - 2

## Data warehouse design

- Top-down approach
  - the data warehouse provides a global and complete representation of business data
  - significant cost and time consuming implementation
  - complex analysis and design tasks
- Bottom-up approach
  - incremental growth of the data warehouse, by adding data marts on specific business areas
  - separately focused on specific business areas
  - limited cost and delivery time
  - easy to perform intermediate checks

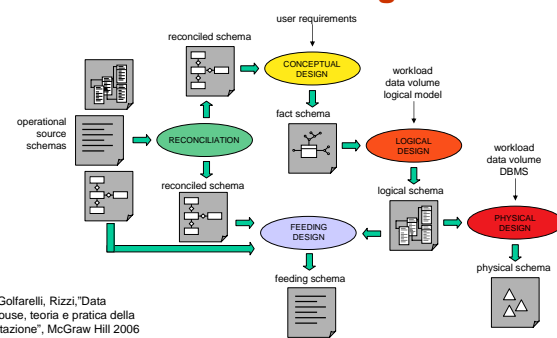
Database and data mining group, Politecnico di Torino  
Copyright - All rights reserved  
DATA WAREHOUSE: DESIGN - 3

## Business Dimensional Lifecycle (Kimball)



From Goffarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006  
Database and data mining group, Politecnico di Torino  
Copyright - All rights reserved  
DATA WAREHOUSE: DESIGN - 4

## Data mart design



From Goffarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006  
Database and data mining group, Politecnico di Torino  
Copyright - All rights reserved  
DATA WAREHOUSE: DESIGN - 5

## Requirement analysis

- It collects
  - data analysis requirements to be supported by the data mart
  - implementation constraints due to existing information systems
- Requirement sources
  - business users
  - operational system administrators
- The first selected data mart is
  - crucial for the company
  - fed by (few) reliable sources

Database and data mining group, Politecnico di Torino  
Copyright - All rights reserved  
DATA WAREHOUSE: DESIGN - 6

## Application requirements

- Description of relevant events (facts)
  - each fact represents a category of events which are relevant for the company
    - examples: (in the CRM domain) complaints, services
  - characterized by descriptive dimensions (setting the granularity), history span, relevant measures
  - informations are gathered in a glossary
- Workload description
  - periodical business reports
  - queries expressed in natural language
    - example: number of complaints for each product in the last month

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 7      Elena Baralis Politecnico di Torino

## Structural requirements

- Feeding periodicity
- Available space for
  - data
  - derived data (indices, materialized views)
- System architecture
  - level number
  - dependent or independent data marts
- Deployment planning
  - start up
  - training

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 8      Elena Baralis Politecnico di Torino

# Conceptual design

Elena Baralis  
Politecnico di Torino

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 9      Elena Baralis Politecnico di Torino

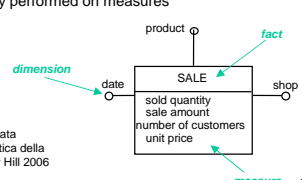
## Conceptual design

- 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

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 10      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

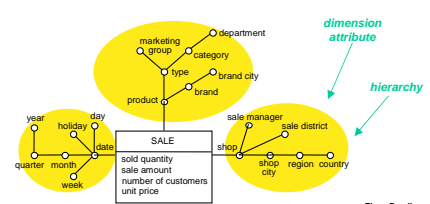


From Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 11      Elena Baralis Politecnico di Torino

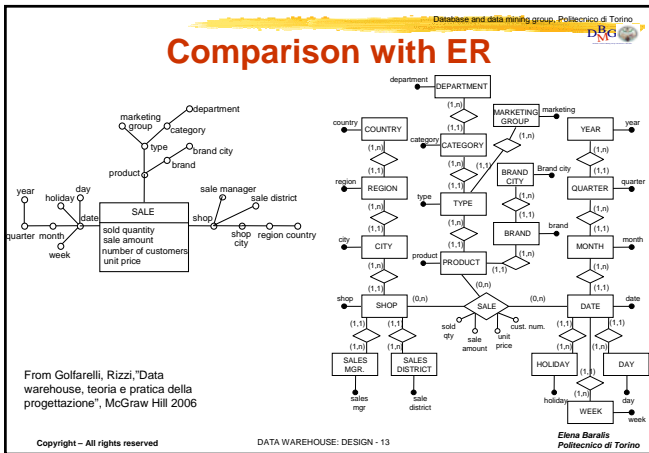
## DFM: Hierarchy

- 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, teoria e pratica della progettazione", McGraw Hill 2006

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 12      Elena Baralis Politecnico di Torino



## Aggregation

- Events can be aggregated based on the values of the attributes along the hierarchies
- Measures for the aggregated event are obtained by aggregating the measures of the corresponding events in the original fact schema
  - standard aggregate operators: SUM, MIN, MAX, AVG, COUNT
- Aggregation computes measures with a coarser granularity than those in the original fact schema
  - detail reduction is usually obtained by climbing a hierarchy

Copyright - All rights reserved DATA WAREHOUSE: DESIGN - 14 Elena Baralis Politecnico di Torino

## Aggregation

category	type	product	1999				2000			
			Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
home cleaning	Washing powder	Brillo	100	90	95	90	80	70	90	85
		Sbianco	20	30	20	10	25	30	25	20
	soap	Lucido	60	50	60	45	40	40	50	40
		Manipulite	15	20	25	30	15	15	20	10
food	milk	Secur	30	35	20	25	30	30	20	15
		Late F Sharp	90	90	85	75	60	80	85	60
	Late U Sharp	60	80	85	60	70	70	75	65	
	Yogurt Sharp	20	30	40	35	30	35	35	20	
	Bevimi	20	10	25	30	35	30	20	10	
	Colissima	50	60	45	40	50	60	45	40	

Measure: sold quantity

category	1999				2000			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
home cleaning	225	225	240	200	190	185	215	170
food	240	270	280	240	245	275	260	195

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

Copyright - All rights reserved DATA WAREHOUSE: DESIGN - 15 Elena Baralis Politecnico di Torino

## Aggregation

category	type	product	1999				2000			
			Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
home cleaning	Washing powder	Brillo	100	90	95	90	80	70	90	85
		Sbianco	20	30	20	10	25	30	25	20
	soap	Lucido	60	50	60	45	40	40	50	40
		Manipulite	15	20	25	30	15	15	20	10
food	milk	Secur	30	35	20	25	30	30	20	15
		Late F Sharp	90	90	85	75	60	80	85	60
	Late U Sharp	60	80	85	60	70	70	75	65	
	Yogurt Sharp	20	30	40	35	30	35	35	20	
	Bevimi	20	10	25	30	35	30	20	10	
	Colissima	50	60	45	40	50	60	45	40	

Measure: sold quantity

category	1999				2000			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
home cleaning	225	225	240	200	190	185	215	170
food	240	270	280	240	245	275	260	195

category	type	1999		2000	
		Q1	Q2	Q1	Q2
home cleaning	Washing p.	670	603	670	603
	soap	200	155	200	155
food	milk	750	685	750	685
	soda	280	200	280	200

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

Copyright - All rights reserved DATA WAREHOUSE: DESIGN - 16 Elena Baralis Politecnico di Torino

## Aggregation

category	type	product	1999				2000			
			Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
home cleaning	Washing powder	Brillo	100	90	95	90	80	70	90	85
		Sbianco	20	30	20	10	25	30	25	20
	soap	Lucido	60	50	60	45	40	40	50	40
		Manipulite	15	20	25	30	15	15	20	10
food	milk	Secur	30	35	20	25	30	30	20	15
		Late F Sharp	90	90	85	75	60	80	85	60
	Late U Sharp	60	80	85	60	70	70	75	65	
	Yogurt Sharp	20	30	40	35	30	35	35	20	
	Bevimi	20	10	25	30	35	30	20	10	
	Colissima	50	60	45	40	50	60	45	40	

Measure: sold quantity

category	1999				2000			
	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4
home clean	225	225	240	200	190	185	215	170
food	240	270	280	240	245	275	260	195

category	type	1999		2000	
		Q1	Q2	Q1	Q2
home cleaning	Washing p.	670	603	670	603
	soap	200	155	200	155
food	milk	750	685	750	685
	soda	280	200	280	200

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

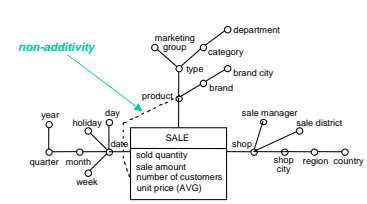
Copyright - All rights reserved DATA WAREHOUSE: DESIGN - 17 Elena Baralis Politecnico di Torino

## Measure characteristics

- additive:
  - can be aggregated along a given hierarchy by means of the SUM operator
- not additive:
  - cannot be aggregated along a given hierarchy by means of the SUM operator
- not aggregable:
  - cannot be aggregated along any hierarchy by means of any aggregate operator

Copyright - All rights reserved DATA WAREHOUSE: DESIGN - 18 Elena Baralis Politecnico di Torino

## Aggregation



Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 19      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

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 20      Elena Baralis Politecnico di Torino

## Aggregate operators

- Distributive
  - can always compute higher level aggregations from more detailed data
  - examples: sum, min, max
- Algebraic
  - can compute higher level aggregations from more detailed data *only* when supplementary support measures are available
  - examples: avg (it requires count)
- Olistic
  - can not compute higher level aggregations from more detailed data
  - examples: mode, median

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 21      Elena Baralis Politecnico di Torino

## Distributive operator: SUM

category	type	product	1999				2000			
			I'99	II'99	III'99	IV'99	I'00	II'00	III'00	IV'00
home cleaning	washing powder	Brillo	100	90	95	90	80	70	90	85
		Sbianco	20	20	20	10	25	20	25	20
		Lucido	60	50	60	45	40	40	50	40
soap		Mangiapelle	15	20	25	30	15	15	20	10
		Scrub	30	25	20	25	30	30	20	15
		Late F-Shap	90	90	85	75	60	80	85	60
food	milk	Late U-Shap	60	80	85	60	70	70	75	65
		Yogurt Shap	20	40	40	35	30	35	35	20
	soda	Bevumi	20	10	25	30	35	30	20	10
	Colissima	50	60	45	40	50	60	45	40	

Measure: sold quantity

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

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 22      Elena Baralis Politecnico di Torino

## Algebraic operator: AVG

category	type	product	1999			
			I'99	II'99	III'99	IV'99
home cleaning	washing powder	Brillo	2	2	2.2	2.5
		Sbianco	1.5	1.5	2	2.5
		Lucido	-	3	3	3
soap		Mangiapelle	1	1.2	1.5	1.5
		Scrub	1.5	1.5	2	-

Measure: unit price

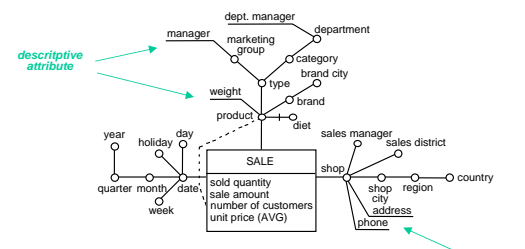
category	type	product	1999			
			I'99	II'99	III'99	IV'99
home cleaning	wash. p.		1.75	2.17	2.40	2.67
		avg:	1.25	1.35	1.25	1.50
soap	avg:		1.50	1.76	2.08	2.09

category	type	product	1999			
			I'99	II'99	III'99	IV'99
home clean.			1.50	1.84	2.14	2.38

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

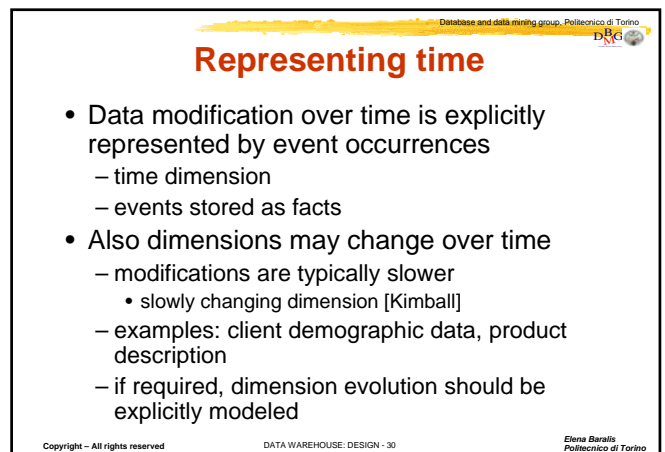
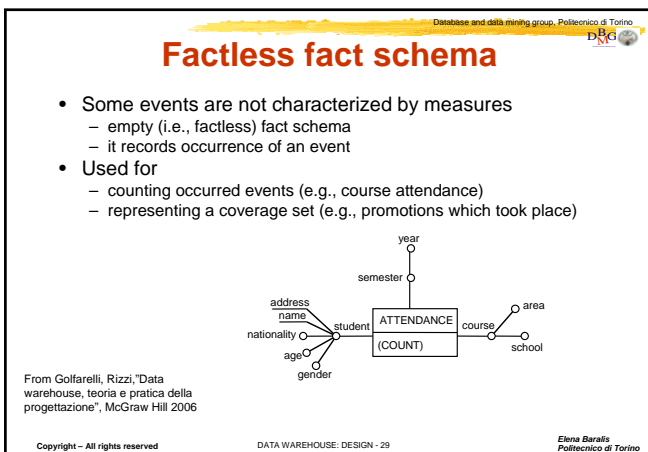
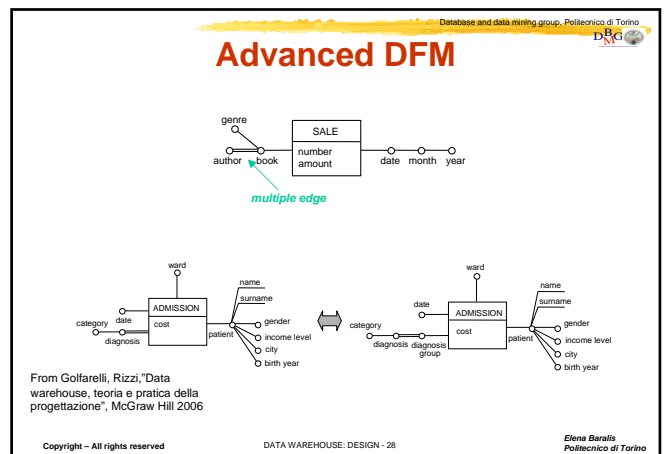
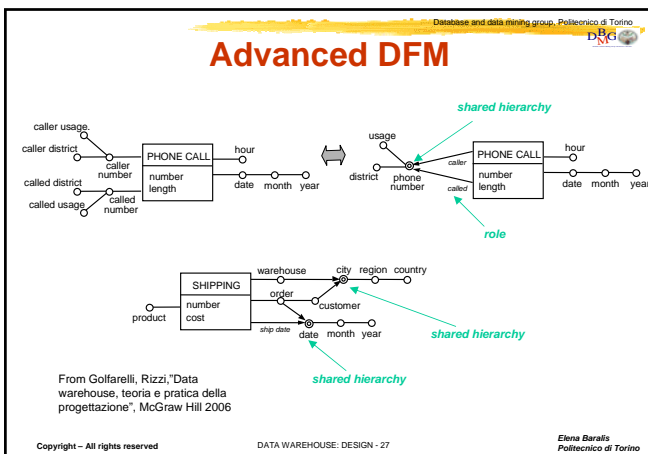
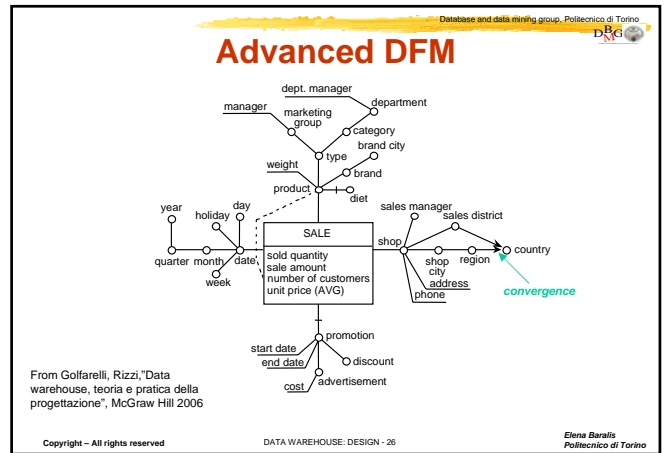
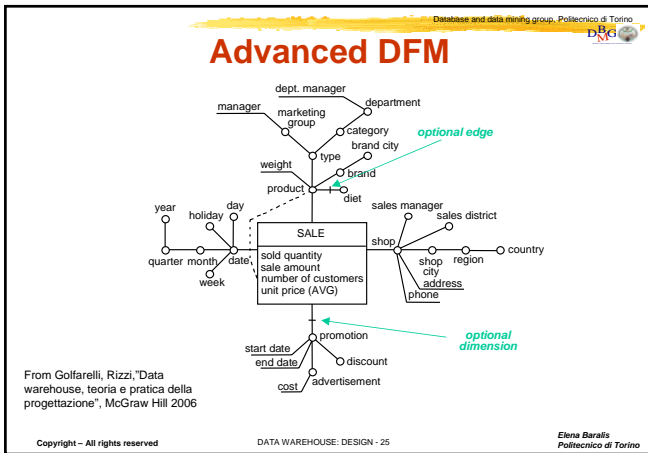
Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 23      Elena Baralis Politecnico di Torino

## Advanced DFM



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

Copyright - All rights reserved      DATA WAREHOUSE: DESIGN - 24      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
- Also known as type I

Copyright – All rights reserved

DATA WAREHOUSE: DESIGN - 31

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)
- Also known as type II

Copyright – All rights reserved

DATA WAREHOUSE: DESIGN - 32

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
  - 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
- Also known as type III

Copyright – All rights reserved

DATA WAREHOUSE: DESIGN - 33

Elena Baralis  
Politecnico di Torino

## Workload

- Workload defined by
  - standard reports
  - approximate estimates discussed with users
- Actual workload difficult to evaluate at design time
  - if the data warehouse succeeds, user and query number may grow
  - query type may vary over time
- Data warehouse tuning
  - performed after system deployment
  - requires monitoring the actual system workload

Copyright – All rights reserved

DATA WAREHOUSE: DESIGN - 34

Elena Baralis  
Politecnico di Torino

## Data volume

- Estimation of the space required by the data mart
  - for data
  - for derived data (indices, materialized views)
- To be considered
  - event cardinality for each fact
  - domain cardinality (number of distinct values) for hierarchy attributes
  - attribute length
- It depends on the temporal span of data storage
- Sparsity
  - occurred events are not all combinations of the dimension elements
  - example: the percentage of products actually sold in each shop and day is roughly 10% of all combinations

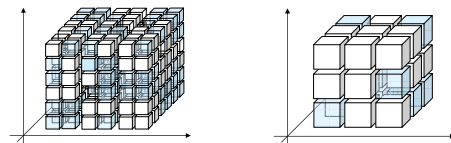
Copyright – All rights reserved

DATA WAREHOUSE: DESIGN - 35

Elena Baralis  
Politecnico di Torino

## Sparsity

- It decreases with increasing data aggregation level
- May significantly affect the accuracy in estimating aggregated data cardinality



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

Copyright – All rights reserved

DATA WAREHOUSE: DESIGN - 36

Elena Baralis  
Politecnico di Torino