UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Data WareHouse – Conceptual modeling

Prof. A. Peron

Slides from M. Golfarelli, S. Rizzi, Datawarehouse Design, Modern Principles and methodologies, McGrawHill.

(Slightly modified by Dario Della Monica)

Dimensional Fact Model

- A conceptual model model created specifically to function as a datamart design support
- It is essentially graphic and based on the multidimensional model
- **Goals**
- Land effective support to conceptual design;
- Create an environment in which user queries may be formulated intuitively;
- Make communication possible between designers and the user with the goal of formalizing requirement specifications;
- Build a stable platform for logical design;
- Provide clear design documentation.

Dimensional Fact Model

- The conceptual representation generated by the DFM consists of a set of fact schemata
- Fact schemata basically model facts, measures, dimensions, and hierarchies.

▶ **Fact**: A fact is a concept relevant to the decision-making process. It typically models a set of events taking place within a company

▶ **Es.** Examples of facts in the commercial domain are sales, shipments, purchases, complains etc.

▶ It is essential that facts have dynamic properties or evolve in some way over time.

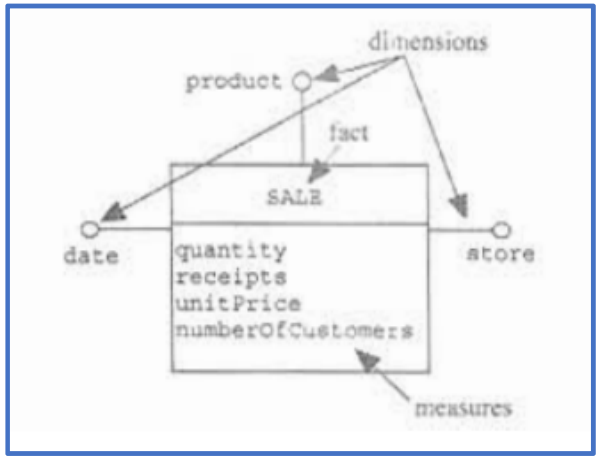
▶ **Measure**: a measure is a numerical property of a fact and describes a quantitative fact aspect that is relevant to analysis

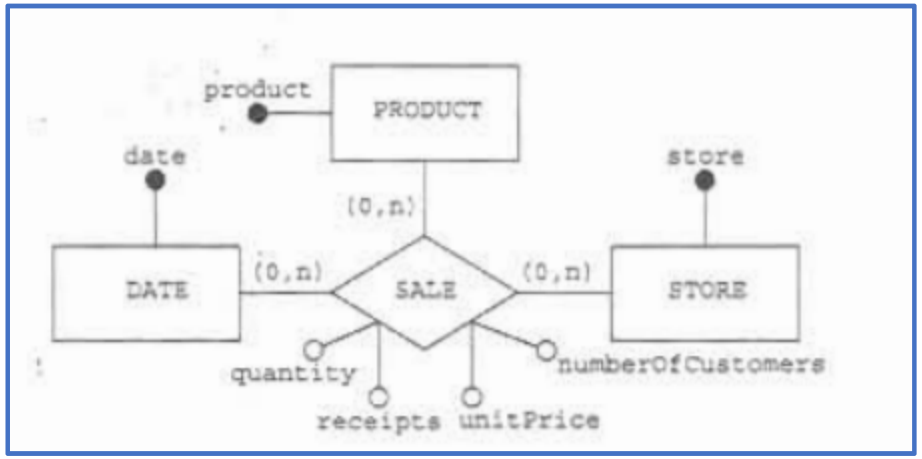
Dimensional Fact Model

- Example of measure. A sale is measured by the number of units sold, the unit price and the total price
- Measures are preferably numeric since they are generally used to make calculations
- A fact may have no measure (one is interested in recording only the occurrence of an event)
- **Dimension:** is a fact property with a finite domain. It describes an analysis coordinate of the fact
- A fact generally has more dimensions that define its finest representation granularity
- Example. Typical dimension of the sales fact are product, store and date; in this case the basic information that can be represented is sales of one product in one store in one day
- After this granularity level it is not possible to distinguish between Sales of a product in a store at different times of day

Dimensional Fact Model

- **Primary event:** a particular occurrence of a fact. It is identified by a one n-ple giving a value for each dimension; a value for each measure is also associated with each primary event
- A graphical representation of a fact schema for sales and the corresponding ER-schema.
- In the relational view a fact is an association among dimensions and measures are attributes of the association.



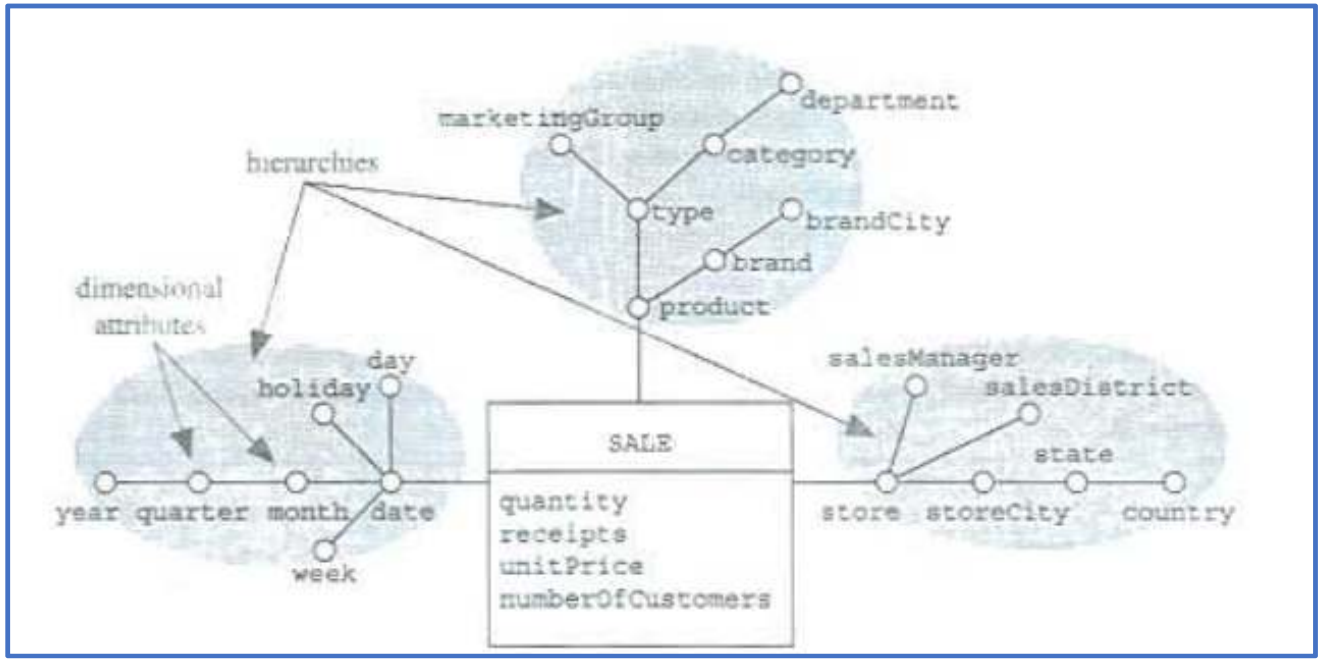


Dimensional Fact Model

- **Dimensional attribute:** attributes with discrete values describing a dimension.
- Example. A product is described by its type, by the category to which it belongs, by its brand and by the department in which it is sold.
- **Hierarchy:** it is (usually) a directed tree:
  - Nodes are dimensional attributes
  - Arcs model (with an exception) 1-to-many associations between pairs of dimensional attributes
  - It includes a dimension attribute positioned at the tree root (finest granularity)
  - It includes all the dimensional attributes of the dimension

Dimensional Fact Model

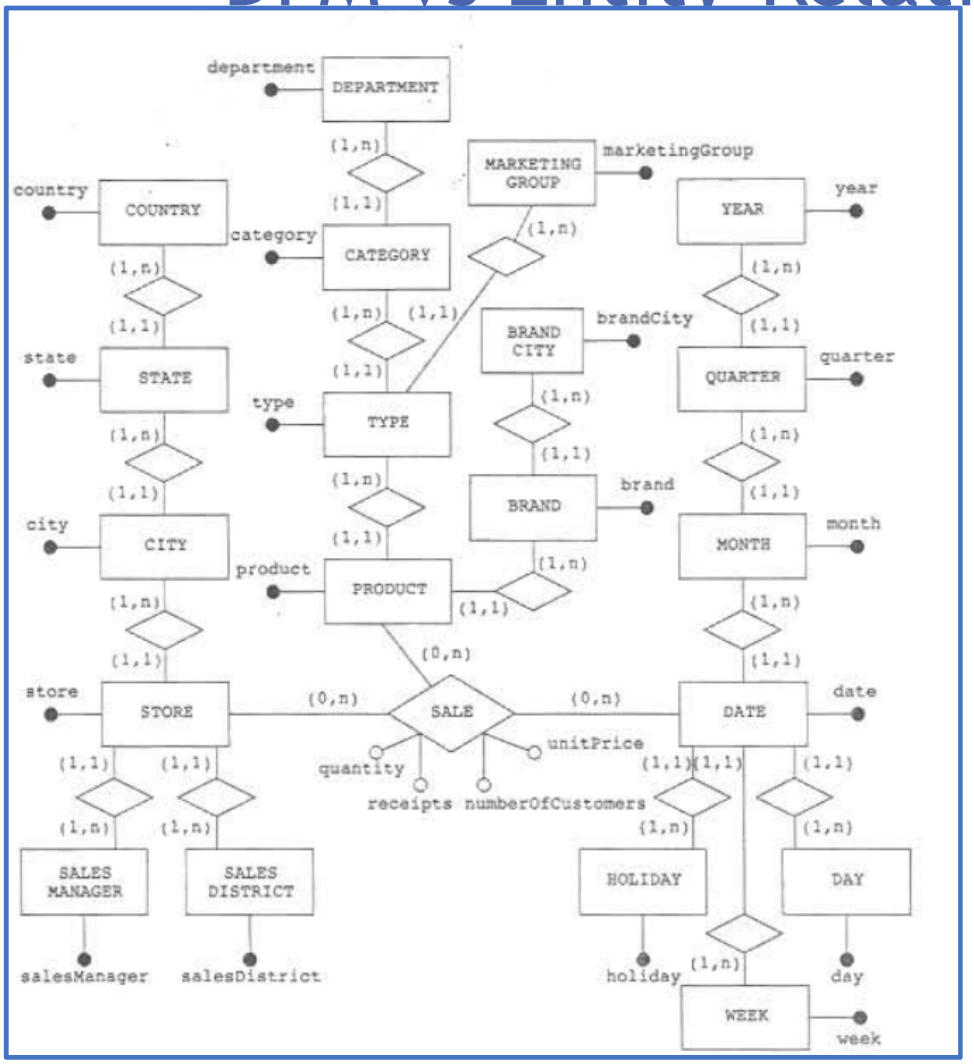
- Since arcs express functional dependences (1-to-many associations) a primary event determines the values of all the dimensional attributes.
- Hierarchies define how primary events can be aggregated and selected for the decisional support;
- The root of a hierarchy provides the more detailed granularity and the other attributes correspond to coarser granularities.

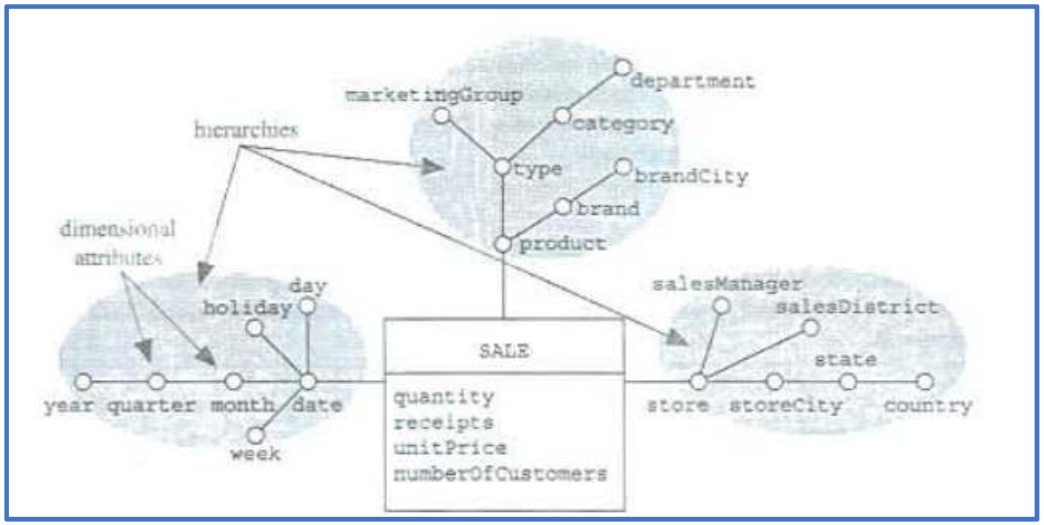


Dimensional Fact Model

- **Secondary event:** Given a set of dimensional attribute covering all dimensions, an assignment of values to them identifies a secondary event that aggregates all of the corresponding primary events. **A secondary event is an instance of a fact (event) at a coarser granularity level.**
- Each secondary event is associated with the value for each measure that aggregates (e.g., sum, average) all the values of the same measure in the corresponding primary events

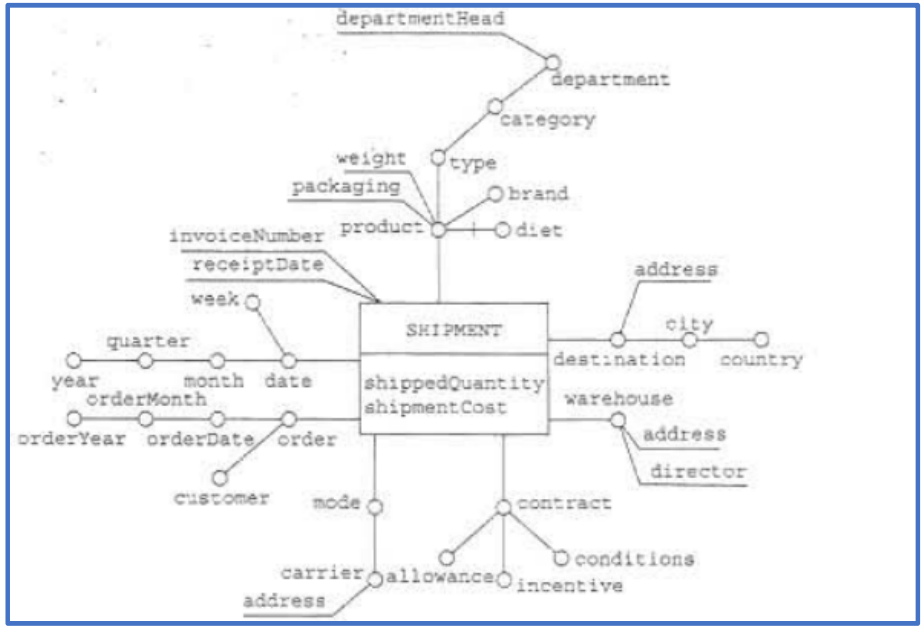
DFM vs Entity-Relationship



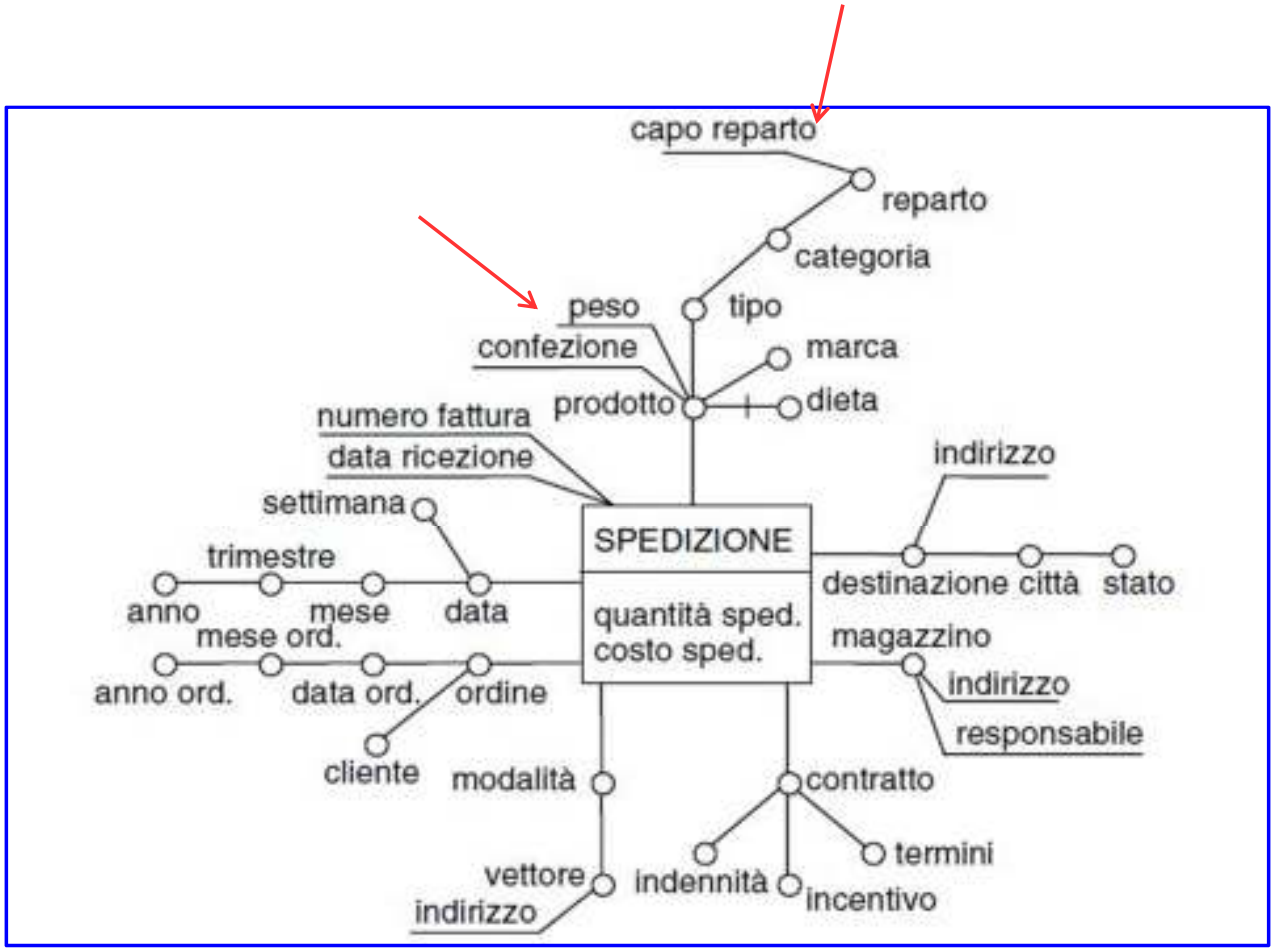


Dimensional Fact Model: enhancements

- **Descriptive attribute:** a property of a dimensional attribute in a hierarchy.
- It is functionally determined by a dimensional attribute.
- It does not add a useful level of aggregation and it is not used for aggregation differently from dimensional attributes.
- Graphically represented in the hierarchy by underlined names.

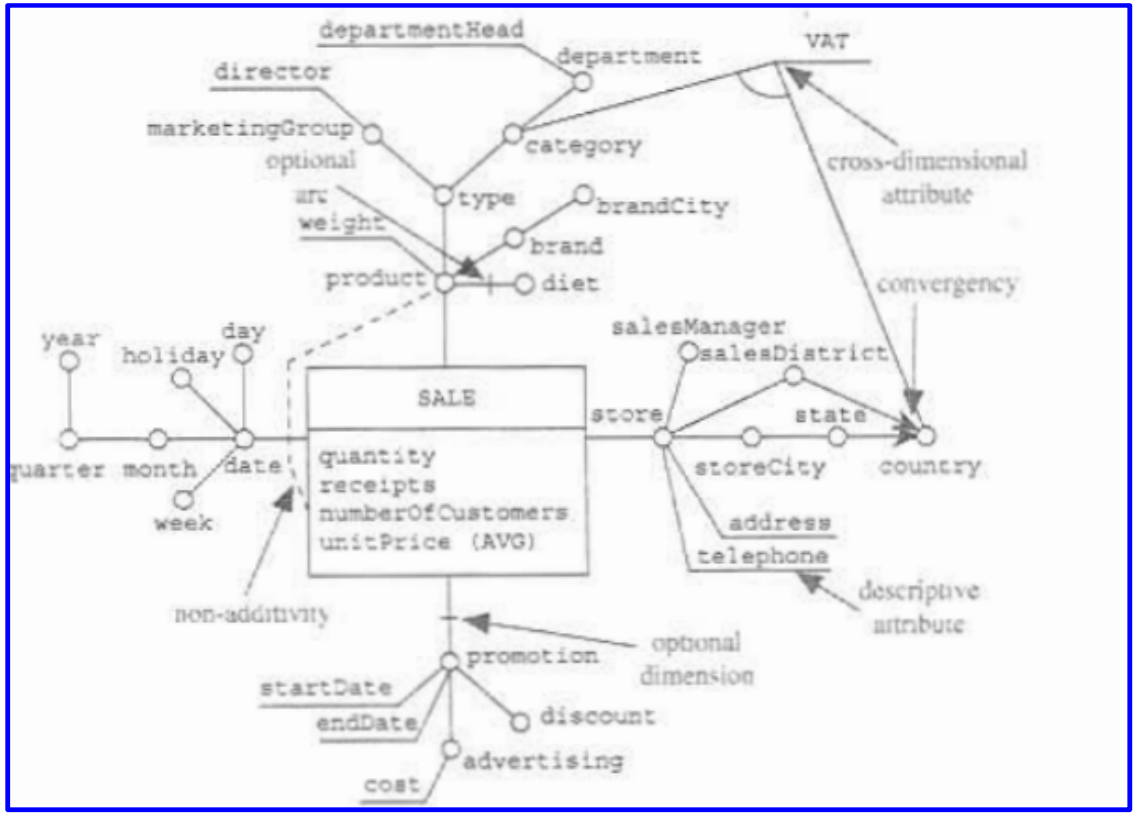


DFM: descriptive attribute



DFM: enhancements (2)

- **Cross-dimensional attributes**: descriptive attribute whose value is defined by the combination of two or more dimensional attributes possibly belonging to different dimensions.
- Graphically represented by joining two arcs with a circular arc.



DFM: enhancements (3)

- **Convergence:** Two or more arcs in the same hierarchy ending to the same dimensional attribute (two fathers for a son in the hierarchy).

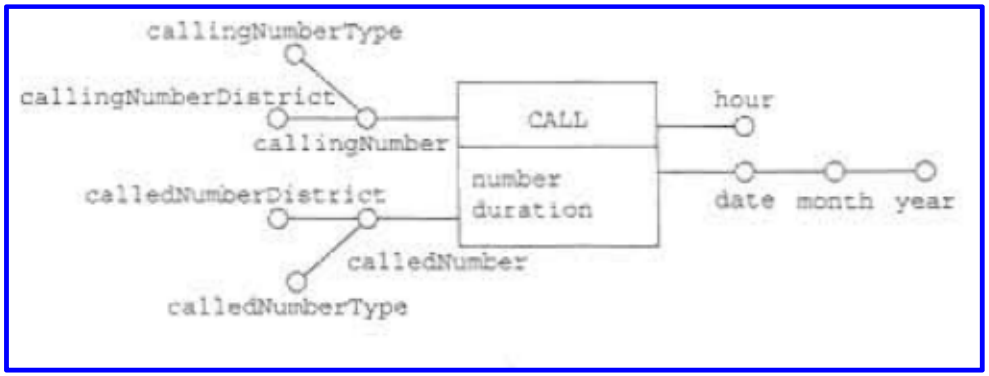
▶ It introduces an exception to the requirement that hierarchies have a tree-like structure .

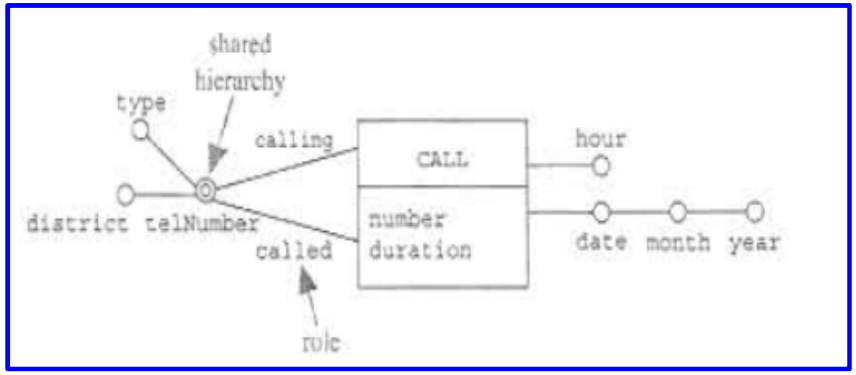
▶Example:

▶Store → storeCity → state → country

▶Store → salesDistrict → country

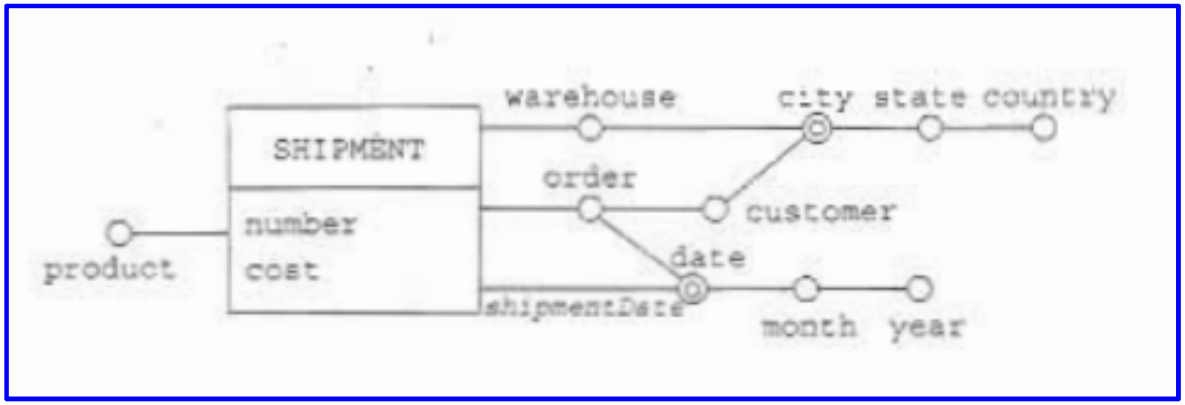
▶ **Shared Hierarchies:** Used to avoid the replication of hierarchies or parts of them.



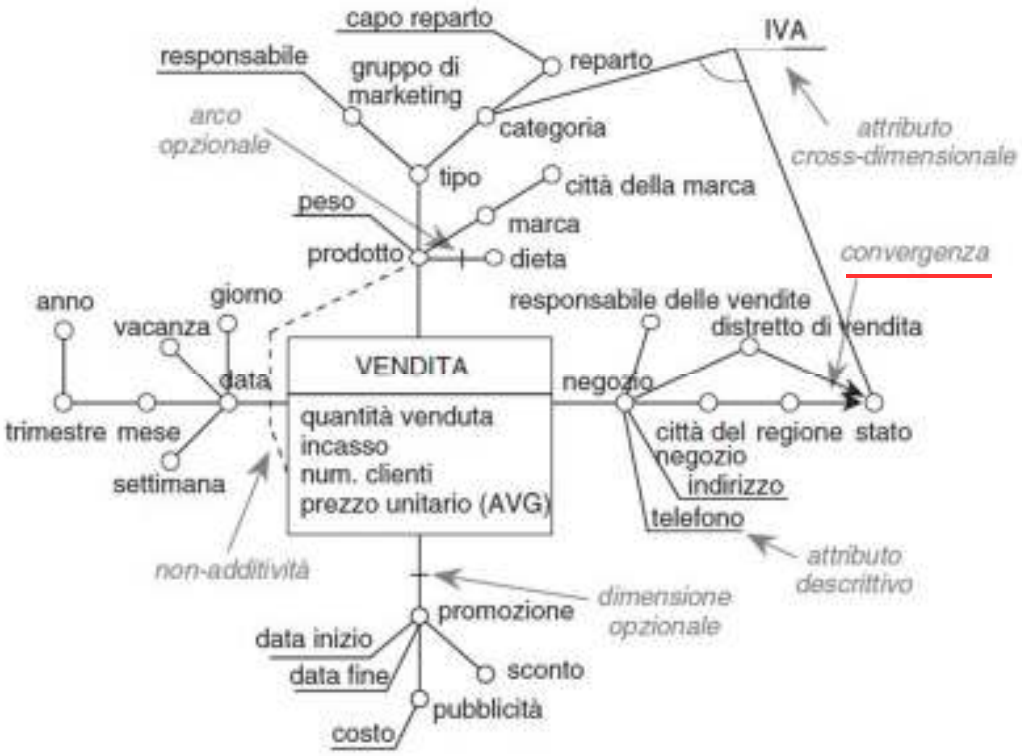


DFM: enhancements (4)

- **Shared Hierarchies:** Used to avoid the replication of hierarchies or parts of them.

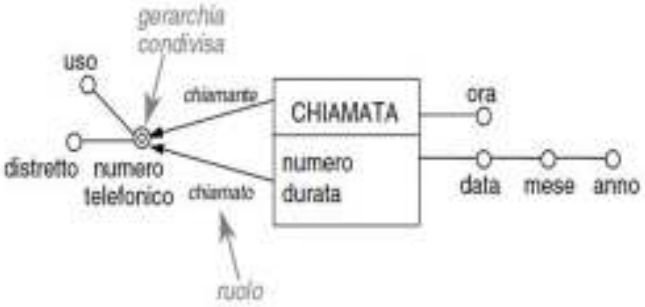


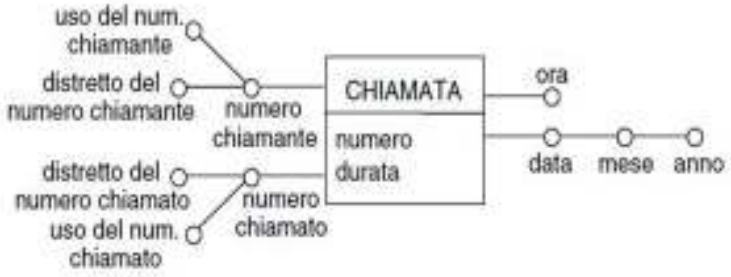
Dimensional Fact Model: estensioni



DFM: estensioni (2)

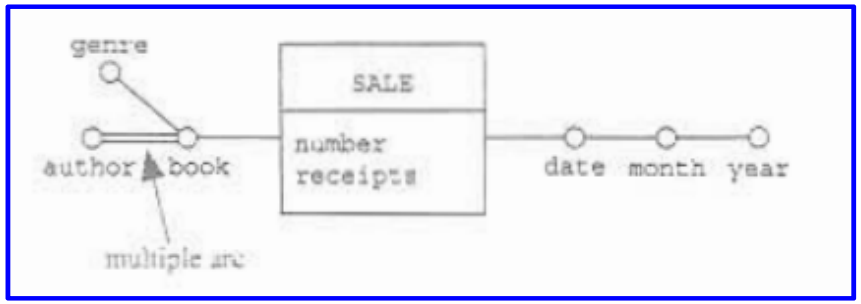
- **Gerarchie condivise**: negli schemi di fatto intere porzioni di gerarchie possono essere replicate due o più volte.
- Es. Gerarchie temporali e gerarchie spaziali. (un fatto può avere come dimensioni più attributi di tipo data o geografici con **significati differenti**)
- L'attributo di partenza per la condivisione viene evidenziato raddoppiando il circolo che lo rappresenta
- tutti i discendenti dell'attributo di partenza sono condivisi.
- Quando la condivisione ha inizio a livello di dimensione, per disambiguare si aggiungere a ciascun arco entrante **un ruolo**.





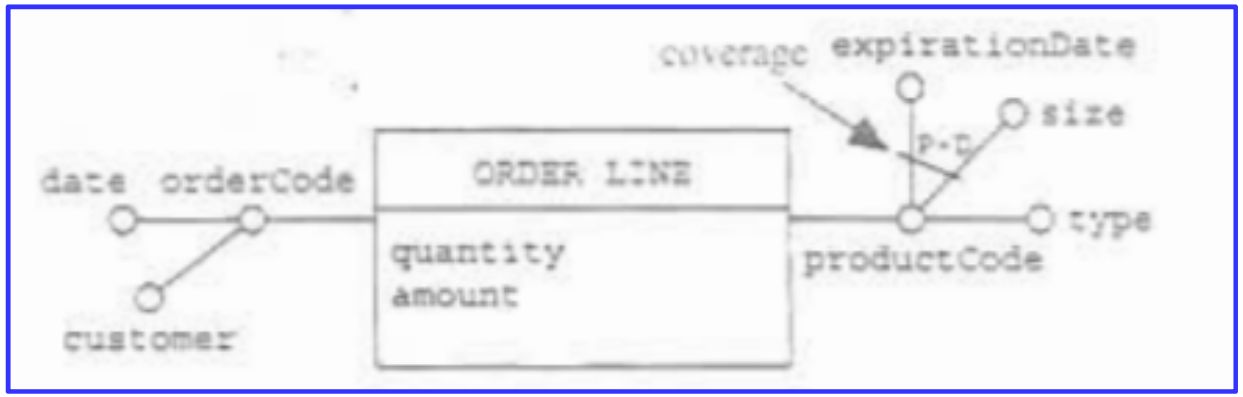
DFM: estensioni (5)

- **Multiple arcs: they are used to describe many-to-many associations.**
- Graphically represented by a double arc.
- **A critical aspect for aggregations!**



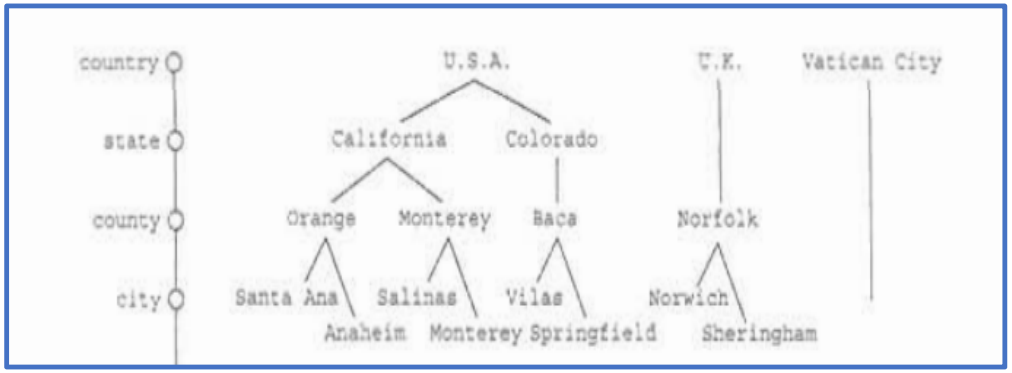
DFM: enhancement (6)

- **Optional arcs:** used for modeling when an association is not total (defined only for a subset of events)
- Graphically represented by an arc marked by a dash.
- When multiple optional arcs exit the same dimensional attribute it is possible to specify the coverage constraints:
- Total/Partial
- Disjoint/Overlapping
- Possible combinations T-D, T-O, P-D e P-O.



DFM: enhancement (6)

- **Incomplete hierarchies**: when for a subset the values of some dimensional attributes are missing (not defined or unknown).
- Some instance may have different levels of aggregation since some level of granularity are missing.
- **Incomplete hierarchies are graphically represented by dashed nodes.**



Dynamic Hierchies

- The standard dynamic aspect is that of fact and their instances (events).
- Also hierchies may have a (slow) dynamic aspect.
- **Examples.**
  - Sales manager rotate among various departments
  - New products can be added (or products can be removed);
  - A product can be moved from a category to another;
  - A store can be moved from a district to another.
- **Structural dynamicity:** involves a change in the structure of a hierarchy and can be handled only with a datamart maintenance step.
- Example:
- Adding a new dimension
- Adding an attribute in a dimension

Dinamicity (2)

Extensional dynamicity: dynamicity in the values which are associated with the hierarchy (the structure does not change)

One has to consider the aspect of extensional dynamicity since they impact the query results.

- When you use a dynamic hierarchy you can actually distinguish four different temporal scenarios in the event analysis:

- **Today-for-Yesterday. All the events are analyzed according to the hierarchy current configuration.**

- **Yesterday-for-Today. All** the events are analyzed according to the configuration the hierarchy had at a previous time.

- **Today or Yesterday. Each event is analyzed according to the configuration the hierarchies had at the time when the event occurred.**

- **Today and Yesterday.** Only the events are considered which refer to the hierarchy instances that remain unchanged

Dinamicity (3)

Example: Assume that the store A changes its name to store B on 1/1/2009.

- **Today-for-Yesterday. All the sales are attributed to store B, even those prior to 1/1/2009**

- **Yesterday-for-Today. All the sales are attributed to store A, even those from 1/1/2009 forward.**

- **Today or Yesterday. The sales prior to 1/1/2009 are attributed to store A and those from 1/1/2009 forward to store B.**

- **Today and Yesterday.** The store sales are not considered since a change has happened.

Additivity

Aggregation requires the definition of a suitable operator to compose the measure values that mark primary events into values to be assigned to secondary events.

Categories of measure:

- **Flow measures**: Refer to a timeframe at the end of which they are evaluated cumulatively.
- Example: the number of products sold in a day, monthly receipts, etc.
- **Level measures**: all evaluated at a particular times.
- Example: the number of products in inventory, the number of inhabitants in a city.
- **Unit measures:** evaluated at particular times but are expressed in relative terms.
- Example: product unit price, discount percentage, currency exchange.

Additivity (2)

Valid aggregation operators for three categories of measures

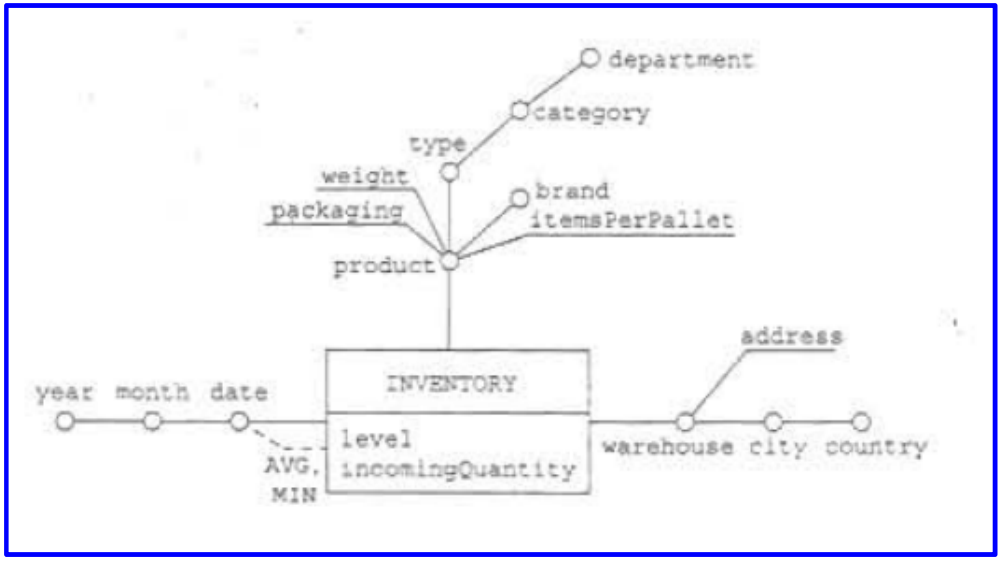
	Temporal Hierarchies	Nontemporal Hierarchies
Flow Measures	SUM, AVG, MIN, MAX	SUM, AVG, MIN, MAX
Level Measures	AVG, MIN, MAX	SUM, AVG, MIN, MAX
Unit Measures	AVG, MIN, MAX	AVG, MIN, MAX

Additivity. The measure is called additive along a dimension when you can use the sum operation to aggregate its values along the dimensional hierarchy. If it is not the case it is called non-additive

A non-addictive measure is non-aggregable when you can use no aggregation operation for it

- Additivity is the most frequent case. So in order to simplify graphic notation in DFM you should explicitly represent only the exceptions.
- A measure is connected to the dimension which is non-additive via a dashed line labeled with the usable aggregation operators.

Additivity (3)



Additivity matrix for the Sales Fact Schema

	product	date	store	promotion
quantity	SUM	SUM	SUM	SUM
receipts	SUM	SUM	SUM	SUM
numberOfCustomers	-	SUM	SUM	SUM
unitPrice	AVG	AVG	AVG	AVG

Agregating additive measures

If a measure is additive with respect to a dimension, the value of a measure M for a secondary event can be calculated by summing the value of M in all the corresponding primary events.

			year	2007				2008			
			quarter	I'07	II'07	III'07	IV'07	I'08	II'08	III'08	IV'08
category	type	product									
House cleaning	Cleaner	Shiny		100	90	95	90	80	70	90	85
		Bleachy		20	30	20	10	25	30	35	20
		Brighty		60	50	60	45	40	40	50	40
	Soap	CleanHand		15	20	25	30	15	15	20	10
		Scent		30	35	20	25	30	30	20	15
Food	Dairy product	F Slurp Milk		90	90	85	75	60	80	85	60
		U Slurp Milk		60	80	85	60	70	70	75	65
		Slurp Yogurt		20	30	40	35	30	35	35	20
	Drink	DrinkMe		20	10	25	30	35	30	20	10
		Coky		50	60	45	40	50	60	45	40

TABLE 5-4 Primary Events of the Sales Cube

	year	2007				2008			
	quarter	I'07	II'07	III'07	IV'07	I'08	II'08	III'08	IV'08
category									
House cleaning		225	225	220	200	190	185	215	170
Food		240	270	280	240	245	275	260	195

TABLE 5-5 The {category, quarter} Group-by Set Secondary Events

Agregating non-additive measures

- For operations different from SUM which can be aggregated along the entire dimension one proceeds exactly as in the previous case: secondary events are calculated from primary events
- Different types of aggregation operations
- Distributive: Calculating aggregates from partial aggregates (e.g. MIN, MAX are distributive)
- Algebraic: Requiring the usage of additional information in the form of support measures to correctly calculate aggregates from partial aggregates (e.g. AVG is algebraic)

Aggregating non-additive measures

			year	2009			
			quarter	I'09	II'09	III'09	IV'09
category	type	product					
House cleaning	Cleaner	Shiny		2	2	2.2	2.5
		Bleachy		1.5	1.5	2	2.5
		Brighty		-	3	3	3
	Soap	CleanHand		1	1.2	1.5	1.5
		Scent		1.5	1.5	2	-

TABLE 5-8 Primary Events of the Sales Cube: A Dash Stands for the Unsold Items in a Quarter

		year	2009			
		quarter	I'09	II'09	III'09	IV'09
category	type					
House cleaning	Cleaner		1.75	2.17	2.40	2.67
	Soap		1.25	1.35	1.75	1.50
	Average:		1.50	1.76	2.08	2.09

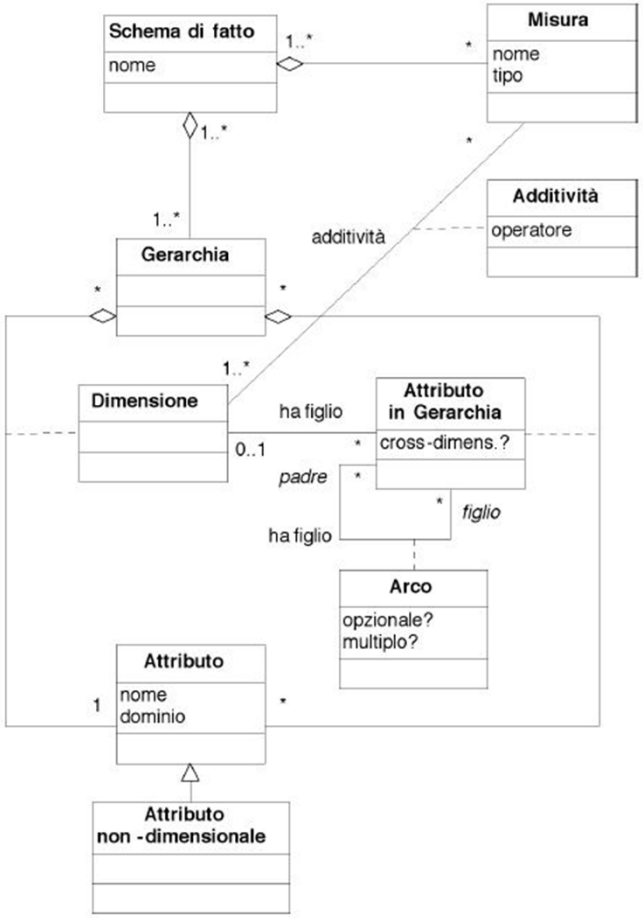
TABLE 5-9 The {type, quarter} Group-by Set Secondary Events

	year	2009			
	quarter	I'09	II'09	III'09	IV'09
category					
House cleaning		1.50	1.84	2.14	2.38

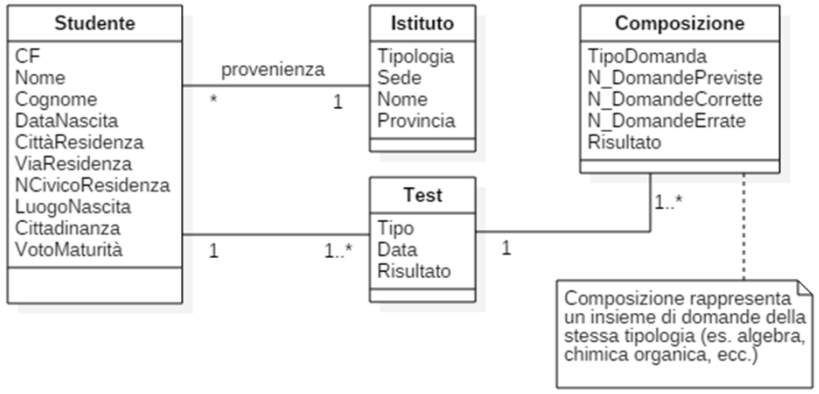
AVG is not distributive

To use aggregated AVG values to obtain the new values, the additional information on the number of items is required

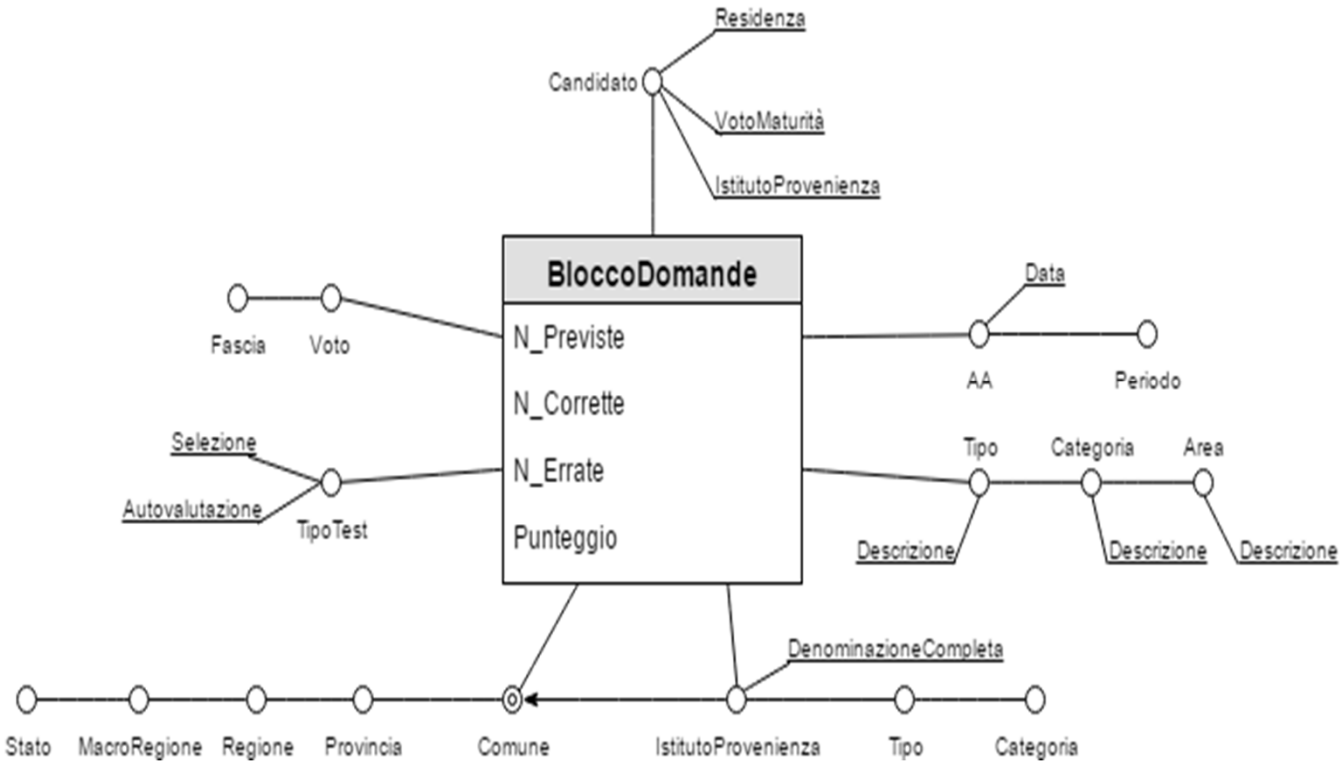
Metamodello DFM



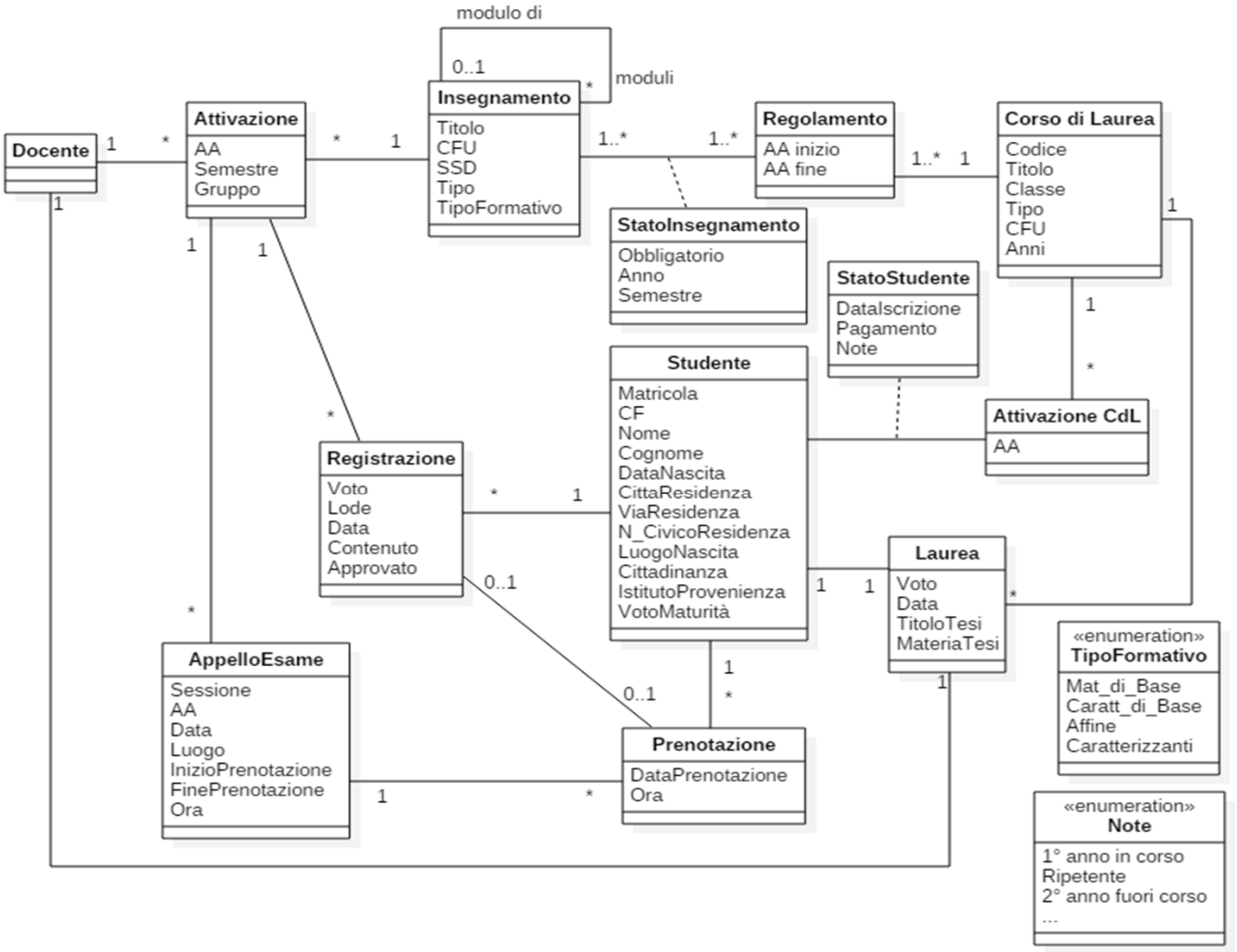
Esempio 1: test ingresso (sorgente oper.)



Esempio 1: test ingresso DFM datamart



Esempio 2: Sistema inf. universitario (sorgente oper.)



Esempio 2: sist. Inf. DFM datamart