Master Degree Programme in Computer Science

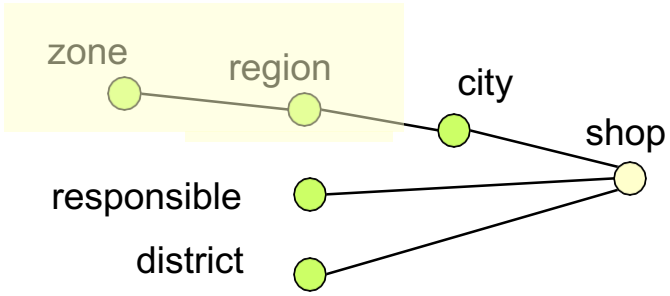
Enterprise Information Systems

SAPIENZA
UNIVERSITÀ DI ROMA

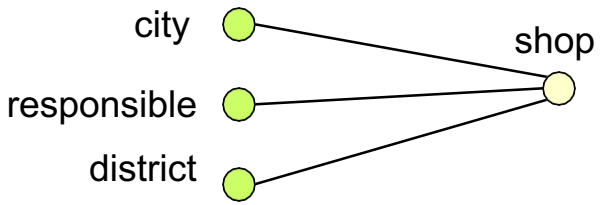
13. Data Warehousing – Logical and Physical Design

Prof. Ing. Claudio CILLI
cilli@di.uniroma1.it
http://dsi.uniroma1.it/~cilli

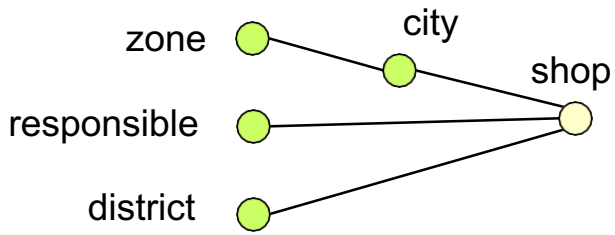
Manipulating hyerarchies



pruning



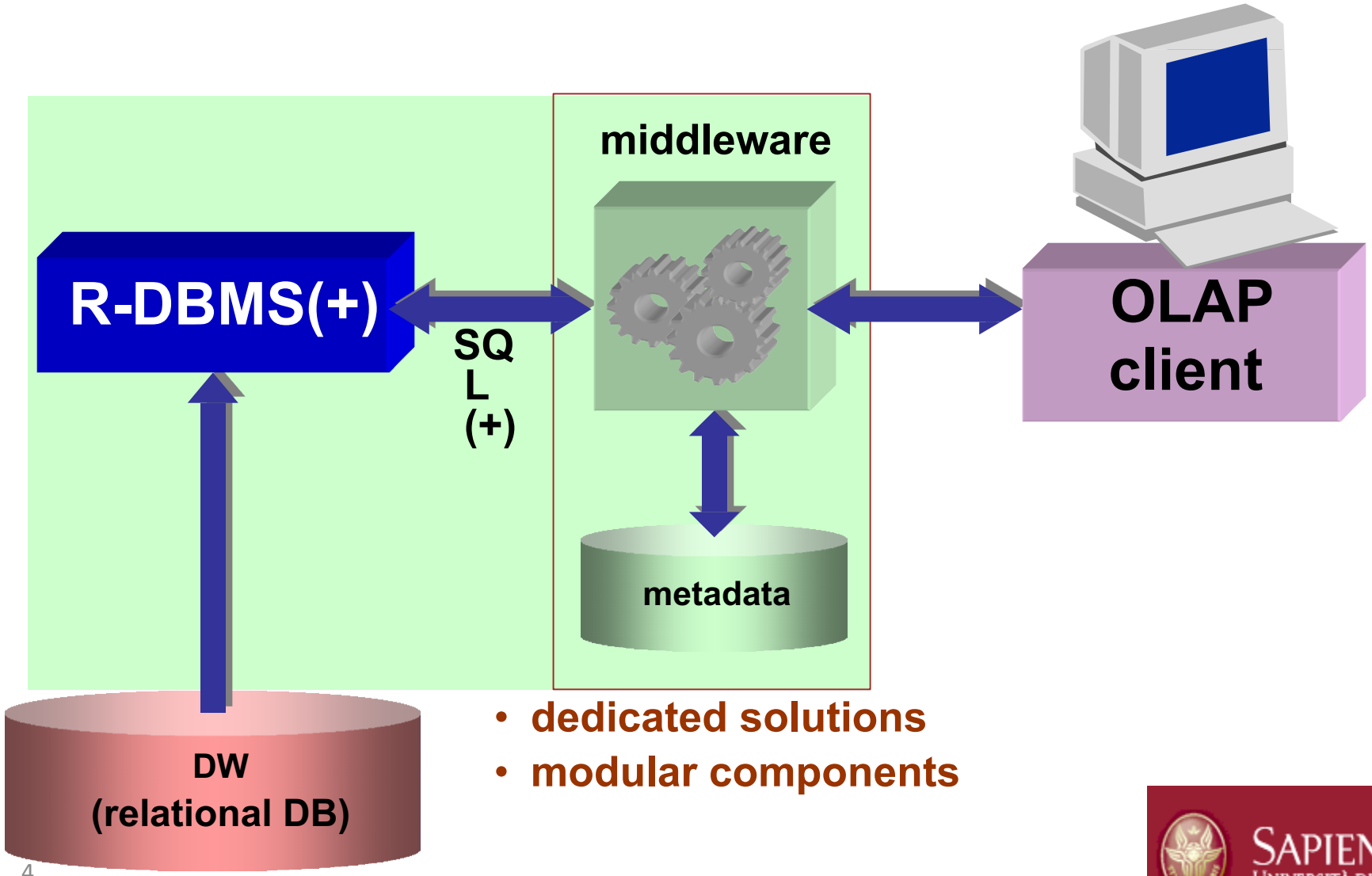
graft



Alternative representations for DW: *OLAP

- **ROLAP** - Relational On-Line Analitical Processing
- data on relational DBMS indexed access
- **MOLAP** - Multidimensional On-Line Analitical Processing
- multidimensional data structures computed access
- **HOLAP** - Hybrid On-Line Analitical Processing
- data on structures of both types introduced by Oracle (Express Server, 2002)

ROLAP Architecture



- dedicated solutions
- modular components

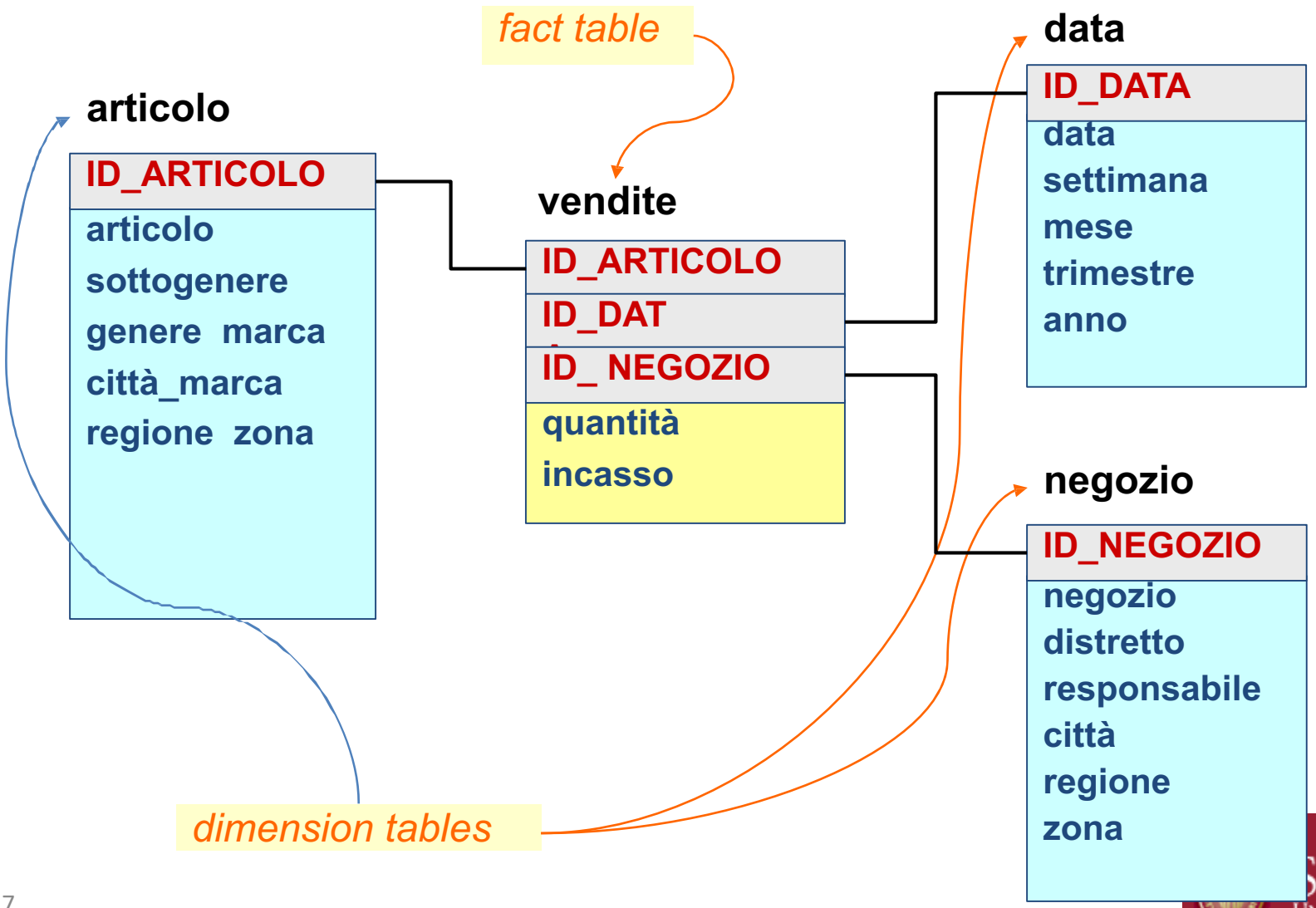
MOLAP Logic Model

- Lack of a standard for both the data structures and the access languages
- Management of the “sparseness” of the data (populated fraction of the multidimensional cube)
  - significant elements identified on the basis of offset (collection of indices of non-zero elements)
  - partitioning into smaller cubes with nearly uniform density (dense or sparse)
  - ad hoc data structures (eg, kd-trees .:)

ROLAP Logic Model: STAR-SCHEMA

- One DIMENSION TABLE for each dimension:
  - primary key (usually a surrogate key)
  - a set of attributes with the values for all levels of aggregation
- A single FACT TABLE:
  - primary key: a foreign-key for each of the dimension tables
  - an attribute for each measure
- Full Denormalization (beside the fact table)

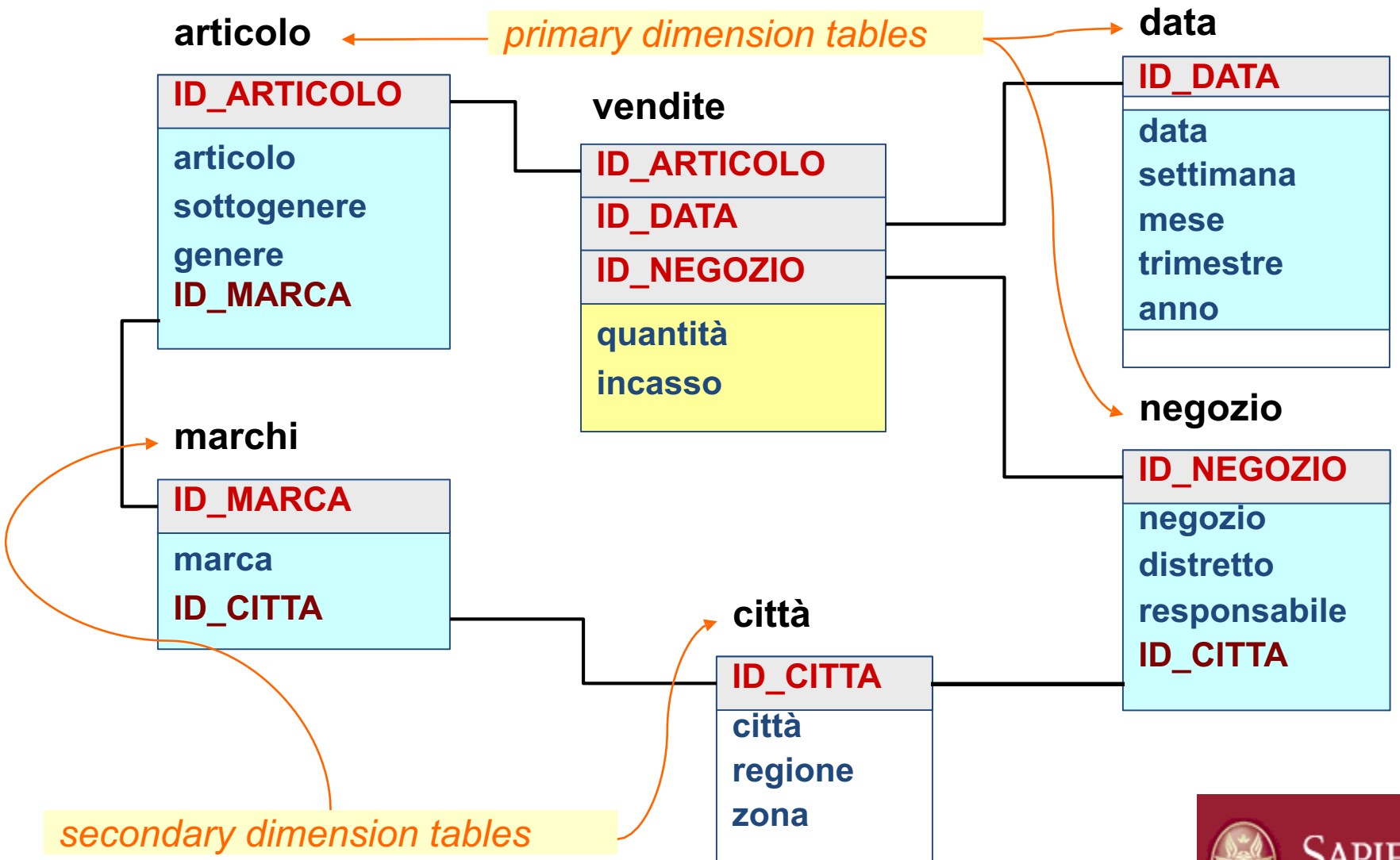
STAR-SCHEMA: example



ROLAP Logic Model: SNOWFLAKE

- Starting from the STAR-SCHEMA, it operates a (partial) NORMALIZATION of the dimension tables, obtaining:
  - for each dimension, the single primary dimension table in the Star-Schema can be decomposed, giving rise to a collection of secondary dimension table
  - a single fact table:
    - primary key, a foreign-key for each of the dimensions (and for each dimension table source)
    - an attribute for each measure

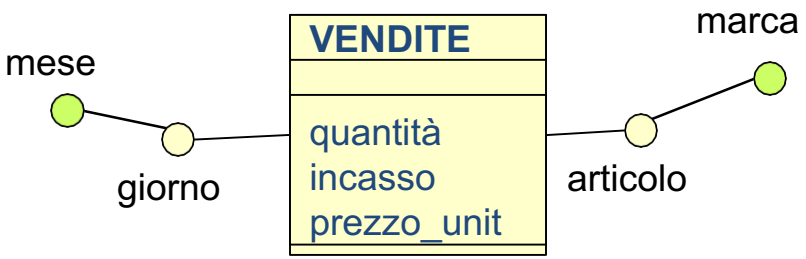
SNOWFLAKE: example

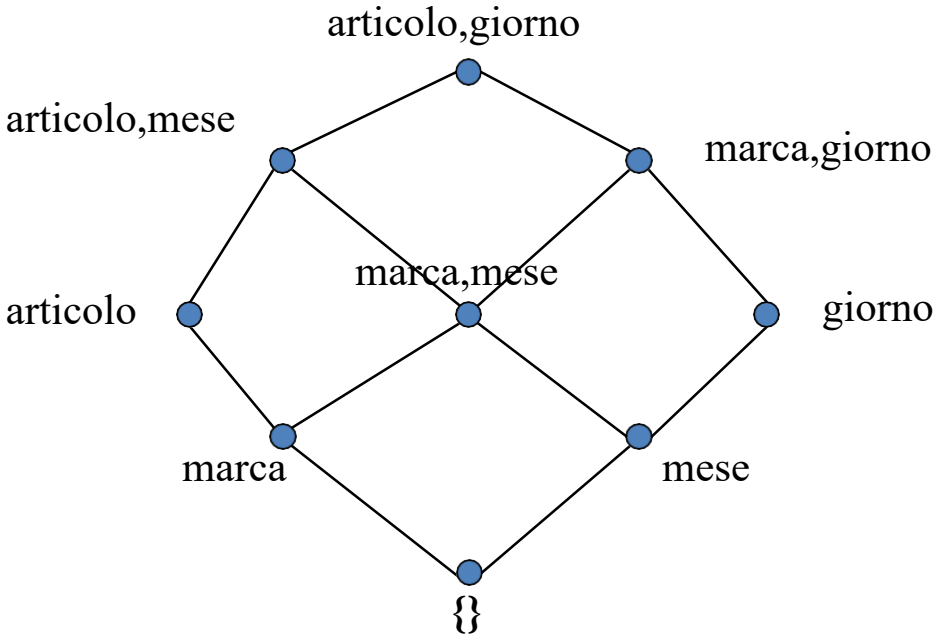


Physical design and VIEWS

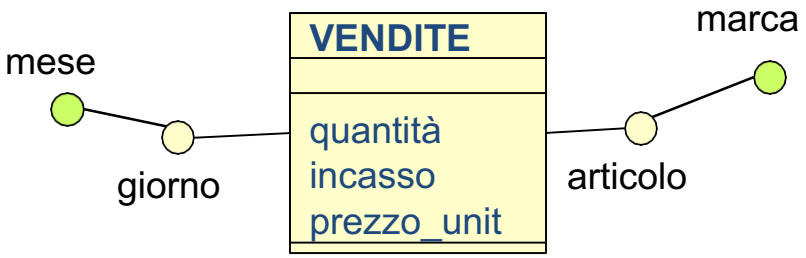
- The main operational issue in a Data Warehouse is the performance
- In contrast, the redundancy is not a serious problem, because of the essentially static data in the DW
- To achieve best performance, it operates a partial materialization of views on the Fact Table
- The compensatory measures related to the materialization of views are:
  - additional space (fully redundant data)
  - calculation time when the refresh of the DW

View (Lattice)





Optimizing computation based on materialized views



COST FACTORS:
- **computing time**
- **space**
- **refresh time**