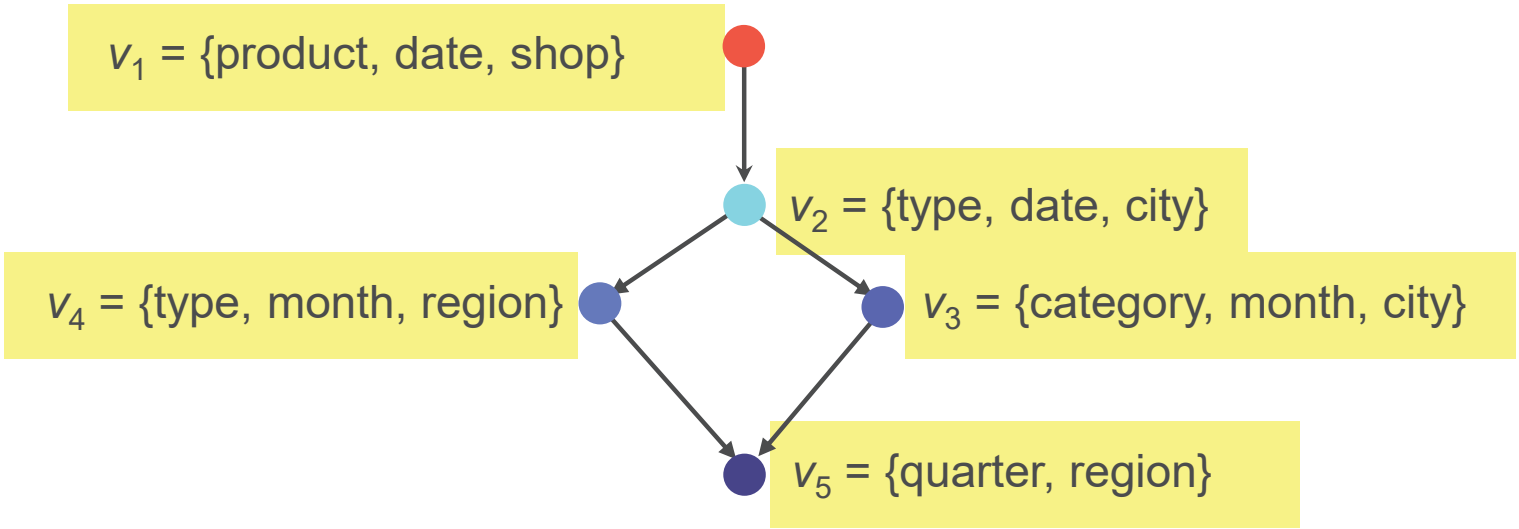
Materialized views

Elena Baralis
Politecnico di Torino

Materialized views

- Precomputed summaries for the fact table
  - explicitly stored in the data warehouse
  - provide a performance increase for aggregate queries



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

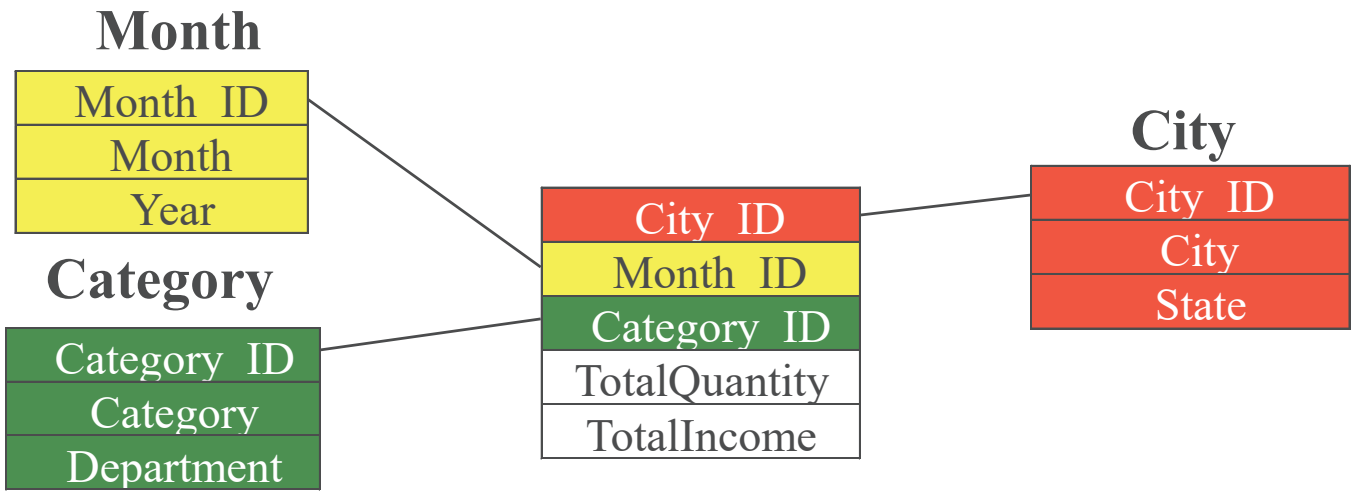
Elena Baralis
Politecnico di Torino

Materialized views

- Defined by SQL statements
- Example: definition of v_3
  - Starting from base tables or views with higher granularity

```
group by City, Category, Month
```

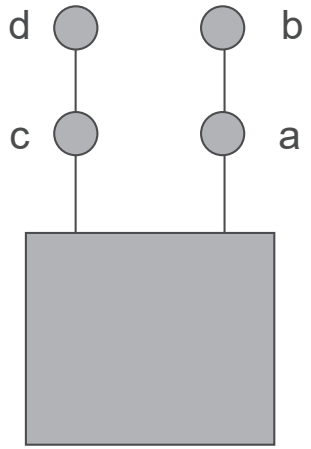
  - Aggregation (SUM) on `Quantity`, `Income` measures
  - Reduction of detail in dimensions

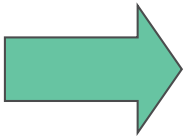


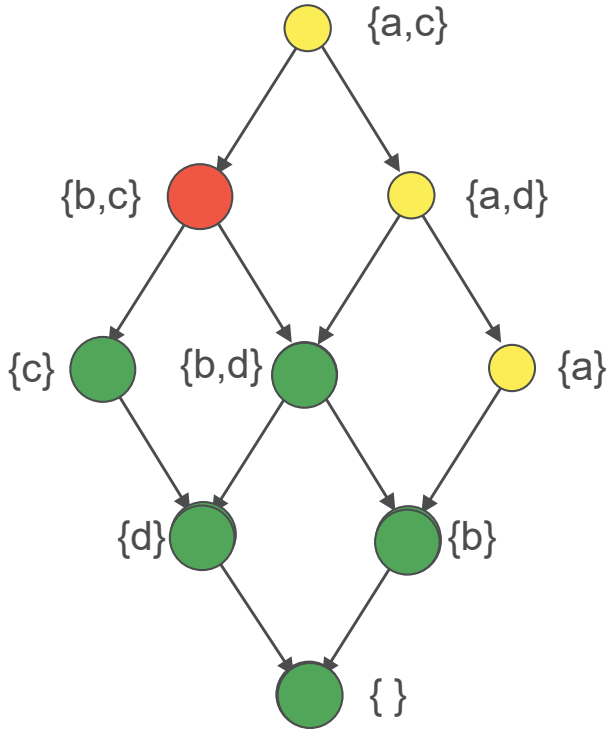
Elena Baralis
Politecnico di Torino

Materialized views

- Materialized views may be exploited for answering several different queries
  - not for all aggregation operators







Multidimensional lattice

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

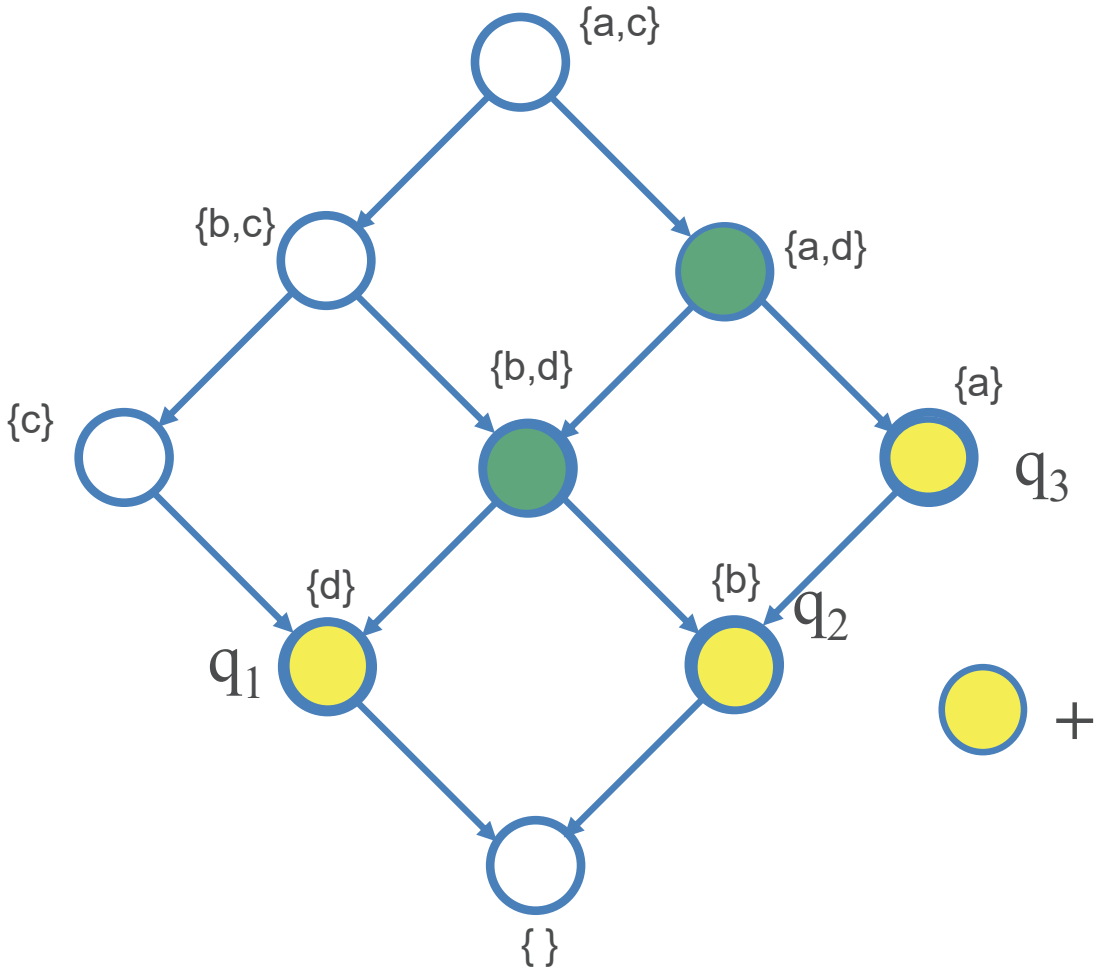
Elena Baralis
Politecnico di Torino

Materialized view selection

- Huge number of allowed aggregations
  - most attribute combinations are eligible
- Selection of the “best” materialized view set
- Cost function minimization
  - query execution cost
  - view maintainance (update) cost
- Constraints
  - available space
  - time window for update
  - response time
  - data freshness

Elena Baralis
Politecnico di Torino

Materialized view selection



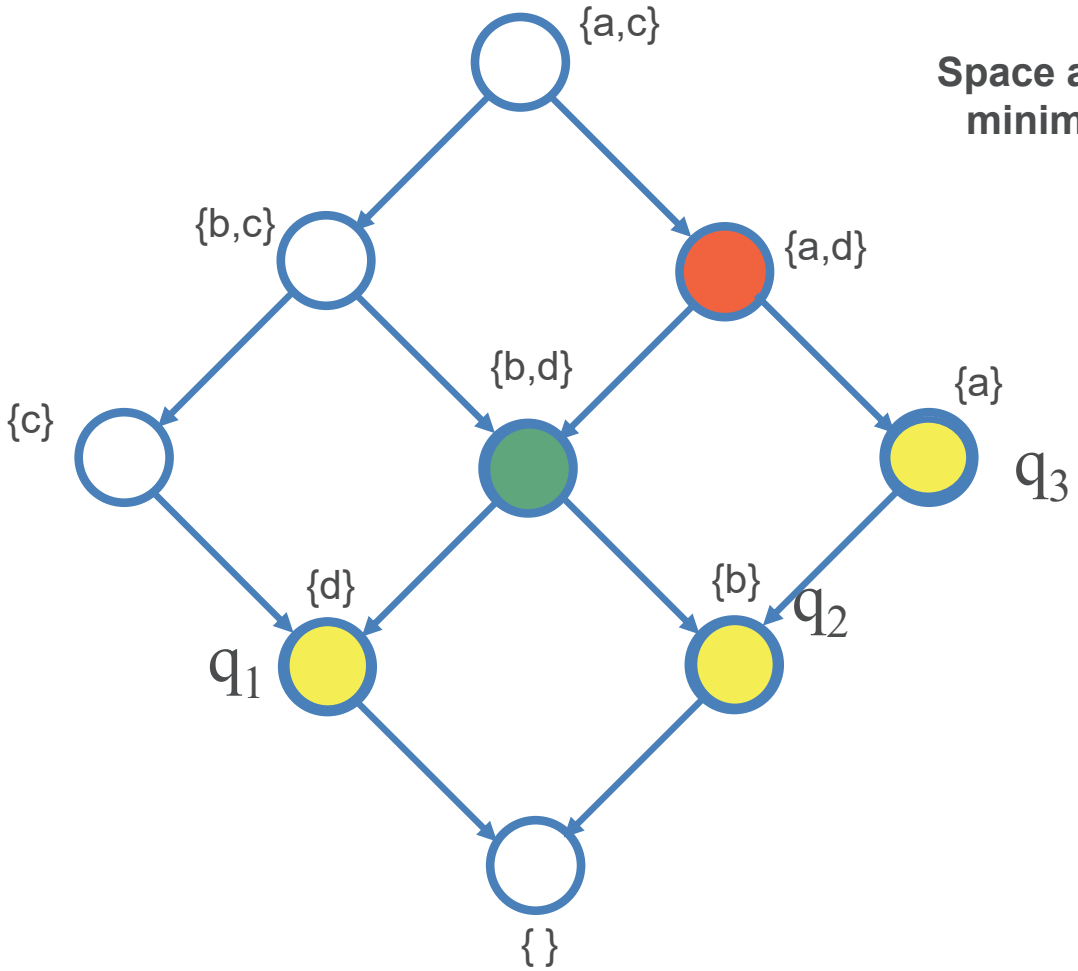
Multidimensional lattice

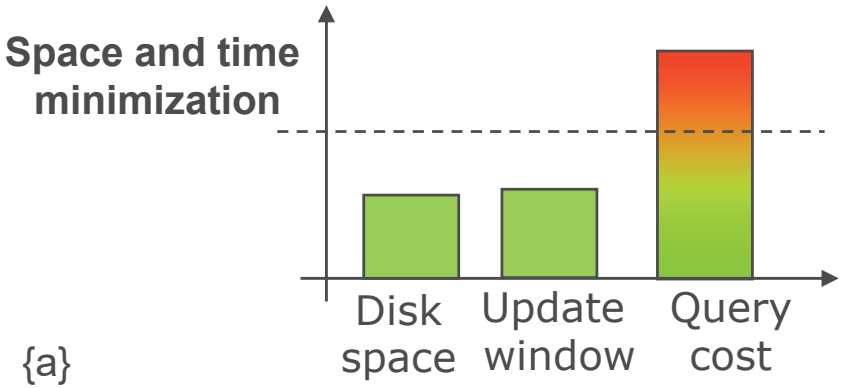
= *candidate views*, **possibly useful to increase workload query performance**

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

Elena Baralis
Politecnico di Torino

Materialized view selection

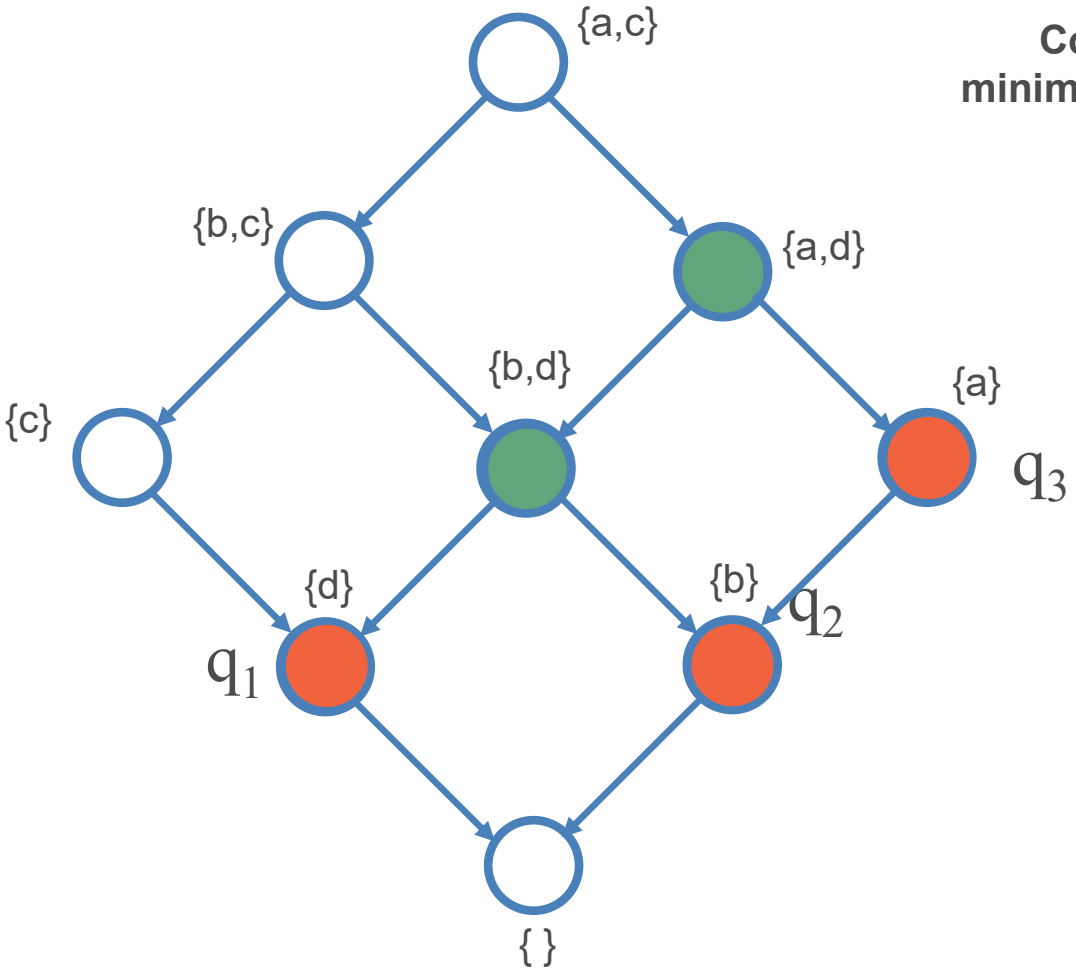


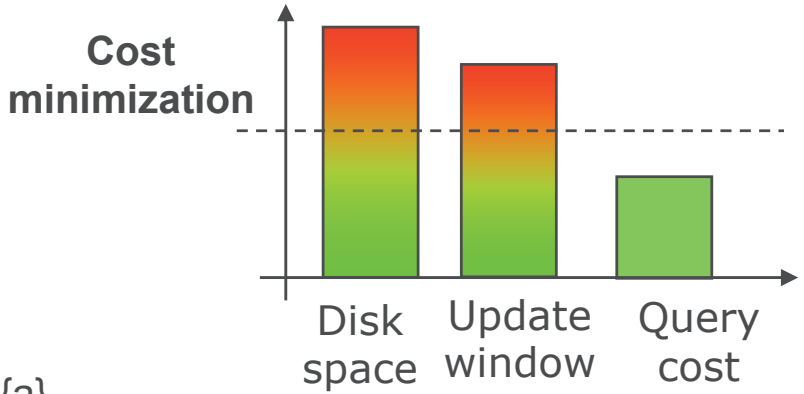


From Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Materialized view selection

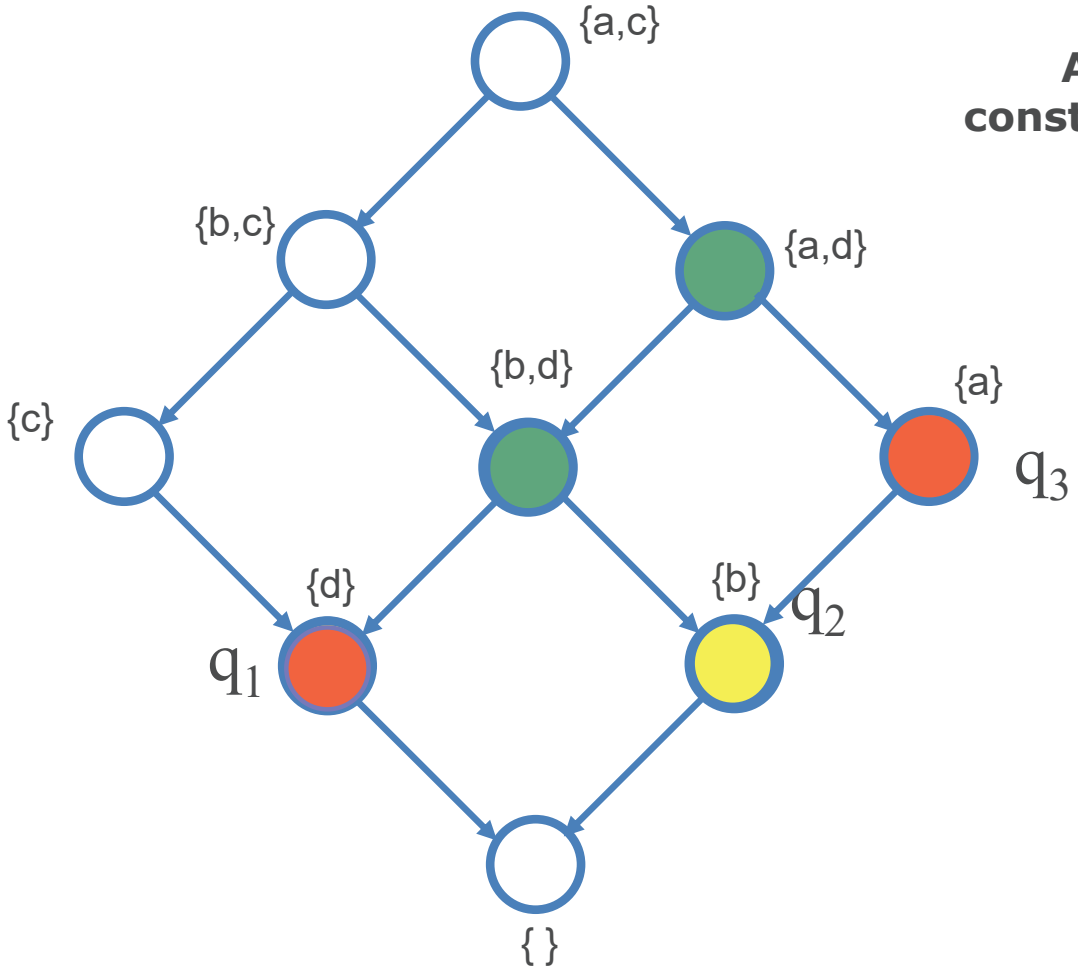


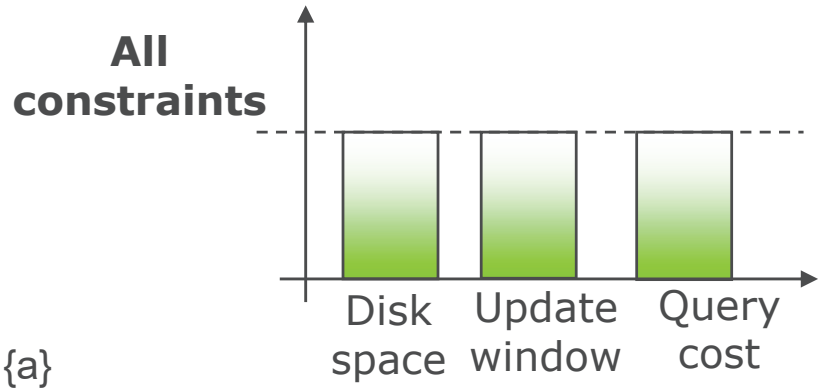


From Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Elena Baralis
Politecnico di Torino

Materialized view selection





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

Elena Baralis
Politecnico di Torino

Physical design

Elena Baralis
Politecnico di Torino

Physical design

- Workload characteristics
  - aggregate queries which require accessing a large fraction of each table
  - read-only access
  - periodic data refresh, possibly rebuilding physical access structures (indices, views)
- Physical structures
  - index types different from OLTP
    - bitmap index, join index, bitmapped join index, ...
    - B^+-tree index not appropriate for
      - attributes with low cardinality domains
      - queries with low selectivity
  - materialized views
    - query optimizer should be able to exploit them

Elena Baralis
Politecnico di Torino

Physical design

- Optimizer characteristics
  - should consider statistics when defining the access plan (cost based)
  - aggregate navigation
- Physical design procedure
  - selection of physical structures supporting most frequent (or most relevant) queries
  - selection of structures improving performance of more than one query
  - constraints
    - disk space
    - available time window for data update

Elena Baralis
Politecnico di Torino

Physical design

- Tuning
  - a posteriori change of physical access structures
  - workload monitoring tools are needed
  - frequently required for OLAP applications
- Parallelism
  - data fragmentation
  - query parallelization
    - inter-query
    - intra-query
  - join and group by lend themselves well to parallel execution

Elena Baralis
Politecnico di Torino

Index selection

- Indexing dimensions
  - attributes frequently involved in selection predicates
  - if domain cardinality is high, then B-tree index
  - if domain cardinality is low, then bitmap index
- Indices for join
  - indexing only foreign keys in the fact table is *rarely* appropriate
  - bitmapped join index is suggested (if available)
- Indices for group by
  - use materialized views

Elena Baralis
Politecnico di Torino

ETL Process

Elena Baralis
Politecnico di Torino

Extraction, Transformation and Loading (ETL)

- Prepares data to be loaded into the data warehouse
  - data extraction from (OLTP and external) sources
  - data cleaning
  - data transformation
  - data loading
- Eased by exploiting the staging area
- Performed
  - when the DW is first loaded
  - during periodical DW refresh

Elena Baralis
Politecnico di Torino

Extraction

- Data acquisition from sources
- Extraction methods
  - static: snapshot of operational data
    - performed during the first DW population
  - incremental: selection of updates that took place after last extraction
    - exploited for periodical DW refresh
    - immediate or deferred
- The selection of which data to extract is based on their quality

Elena Baralis
Politecnico di Torino

Extraction

- It depends on how operational data is collected
  - historical: all modifications are stored for a given time in the OLTP system
    - bank transactions, insurance data
    - operationally simple
  - partly historical: only a limited number of states is stored in the OLTP system
    - operationally complex
  - transient: the OLTP system only keeps the *current* data state
    - example: stock inventory
    - operationally complex

Elena Baralis
Politecnico di Torino

Incremental extraction

- Application assisted
  - data modifications are captured by ad hoc application functions
  - requires changing OLTP applications (or APIs for database access)
  - increases application load
  - hardly avoidable in legacy systems
- Log based
  - log data is accessed by means of appropriate APIs
  - log data format is usually proprietary
  - efficient, no interference with application load

Elena Baralis
Politecnico di Torino

Incremental extraction

- Trigger based
  - triggers capture interesting data modifications
  - does not require changing OLTP applications
  - increases application load
- Timestamp based
  - modified records are marked by the (last) modification timestamp
  - requires modifying the OLTP database schema (and applications)
  - deferred extraction, may lose intermediate states if data is transient

Elena Baralis
Politecnico di Torino

Comparison of extraction techniques

	Static	*Timestamps*	*Appilcation assisted*	*Trigger*	*Log*
Management of transient or semi-periodic data	No	Incomplete	Complete	Complete	Complete
Support to file-based systems	Yes	Yes	Yes	No	Rare
Implementation technique	Tools	Tools or internal developments	Internal developments	Tools	Tools
Costs of enterprise specific development	None	Medium	High	None	None
Use with legacy systems	Yes	Difficult	Difficult	Difficult	Yes
Changes to applications	None	Likely	Likely	None	None
DBMS-dependent procedures	Limited	Limited	Variabile	High	Limited
Impact on operational system performance	None	None	Medium	Medium	None
Complexity of extraction procedures	Low	Low	High	Medium	Low

From Devlin, Data warehouse: from architecture to implementation, Addisono-Wesley, 1997

Elena Baralis
Politecnico di Torino

Incremental extraction

4/4/2010

Cod	Product	Customer	Qty
1	Greco di tufo	Malavasi	50
2	Barolo	Maio	150
3	Barbera	Lumini	75
4	Sangiovese	Cappelli	45

6/4/2010

Cod	Product	Customer	Qty
1	Greco di tufo	Malavasi	50
2	Barolo	Maio	150
4	Sangiovese	Cappelli	145
5	Vermentino	Maltoni	25
6	Trebbiano	Maltoni	150

Incremental difference

Cod	Product	Customer	Qty	Action
3	Barbera	Lumini	75	D
4	Sangiovese	Cappelli	145	U
5	Vermentino	Maltoni	25	I
6	Trebbiano	Maltoni	150	I

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

Elena Baralis
Politecnico di Torino

Data cleaning

- Techniques for improving data quality (correctness and consistency)
  - duplicate data
  - missing data
  - unexpected use of a field
  - impossible or wrong data values
  - inconsistency between logically connected data
- Problems due to
  - data entry errors
  - different field formats
  - evolving business practices

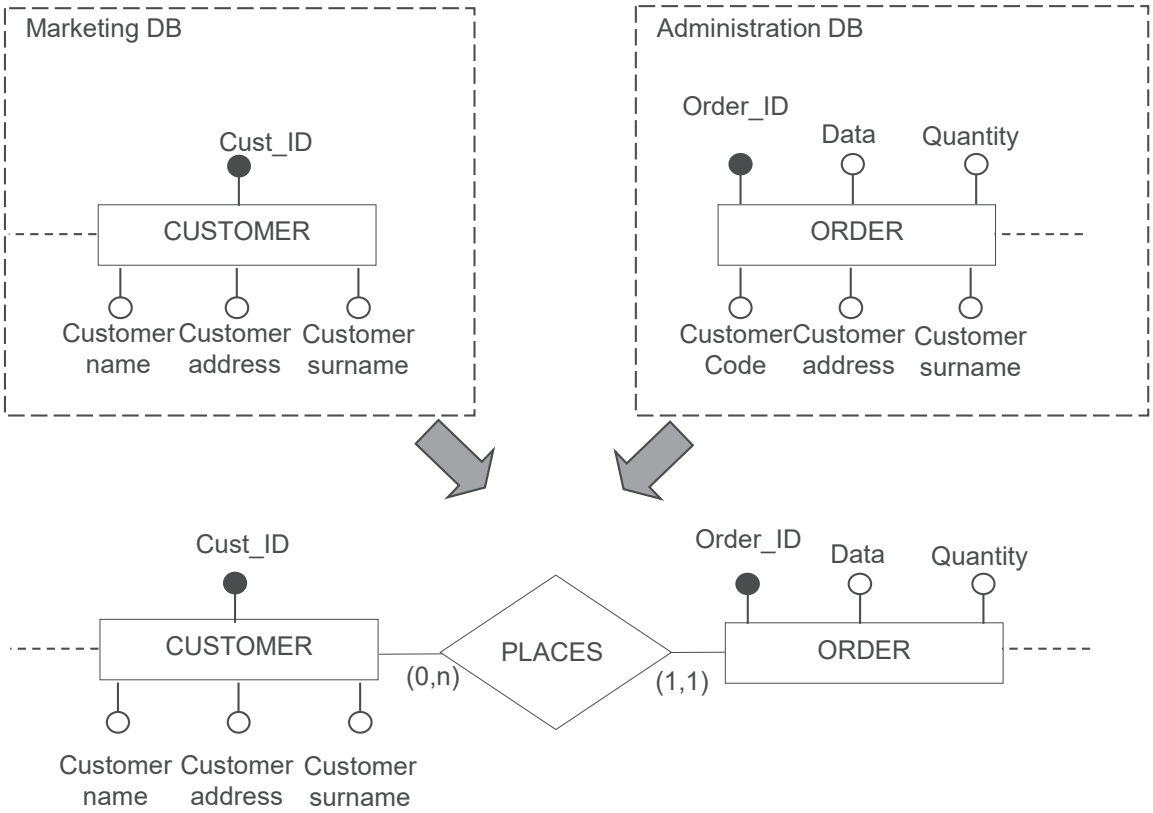
Elena Baralis
Politecnico di Torino

Data cleaning

- Each problem is solved by an ad hoc technique
  - data dictionary
    - appropriate for data entry errors or format errors
    - can be exploited only for data domains with limited cardinality
  - approximate fusion
    - appropriate for detecting duplicates/similar data correlations
      - approximate join
      - purge/merge problem
  - outlier identification, deviations from business rules
- Prevention is the best strategy
  - reliable and rigorous OLTP data entry procedures

Elena Baralis
Politecnico di Torino

Approximate join

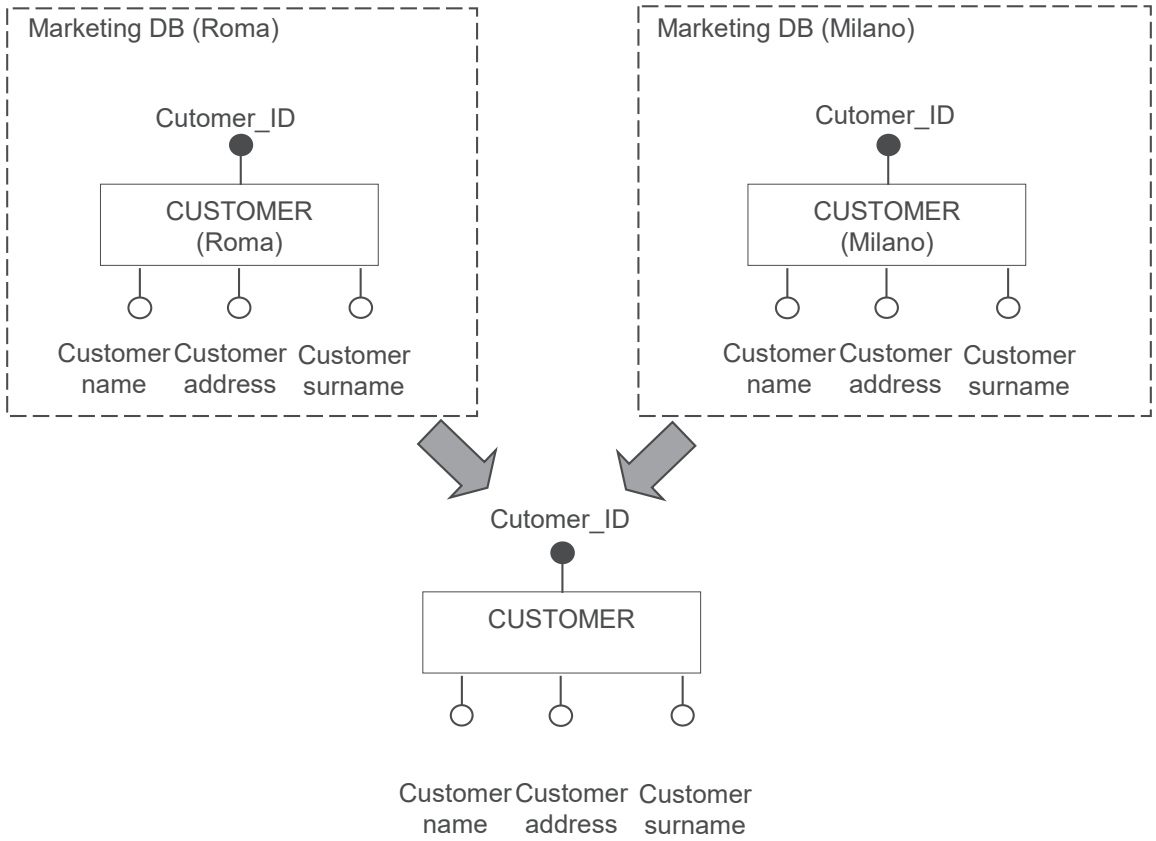


- The join operation should be executed based on common fields, not representing the customer identifier

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

Elena Baralis
Politecnico di Torino

Purge/Merge problem



- Duplicate tuples should be identified and removed
- A criterion is needed to evaluate record similarity

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

Elena Baralis
Politecnico di Torino

Data cleaning and transformation example

Elena Baralis
C.so Duca degli Abruzzi 24
20129 Torino (I)

Normalization

name:	Elena
surname:	Baralis
address:	C.so Duca degli Abruzzi 24
ZIP:	20129
city:	Torino
country:	I

Standardization

name:	Elena
surname:	Baralis
address:	Corso Duca degli Abruzzi 24
ZIP:	20129
city:	Torino
country:	Italia

Correction

name:	Elena
surname:	Baralis
address:	Corso Duca degli Abruzzi 24
ZIP:	10129
city:	Torino
country:	Italia

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

Transformation

- Data conversion from operational format to data warehouse format
  - requires data integration
- A uniform operational data representation (reconciled schema) is needed
- Two steps
  - from operational sources to reconciled data in the staging area
    - conversion and normalization
    - matching
    - (possibly) significant data selection
  - from reconciled data to the data warehouse
    - surrogate keys generation
    - aggregation computation

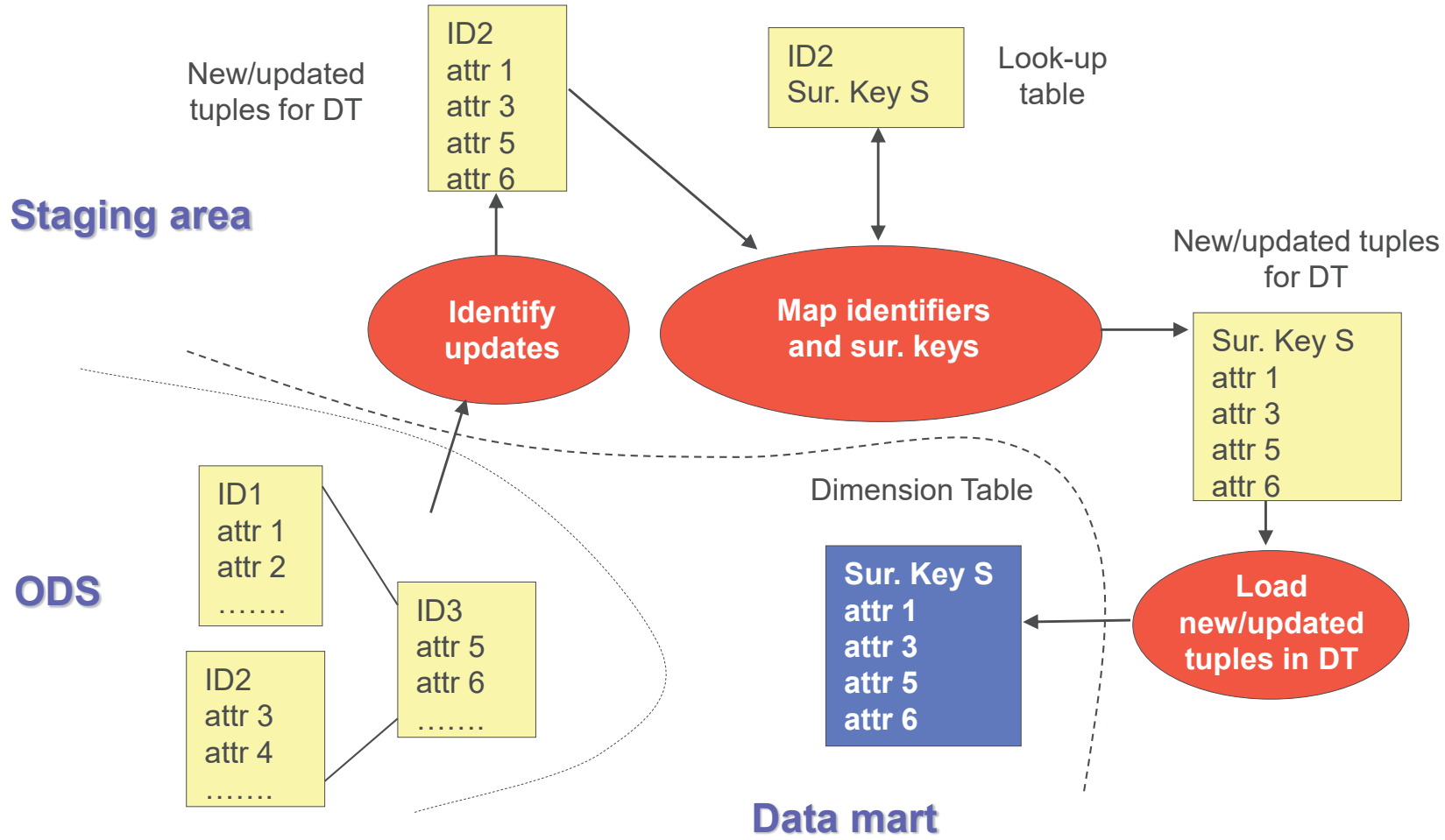
Elena Baralis
Politecnico di Torino

Data warehouse loading

- Update propagation to the data warehouse
- Update order that preserves data integrity
  1. dimensions
  2. fact tables
  3. materialized views and indices
- Limited time window to perform updates
- Transactional properties are needed
  - reliability
  - atomicity

Elena Baralis
Politecnico di Torino

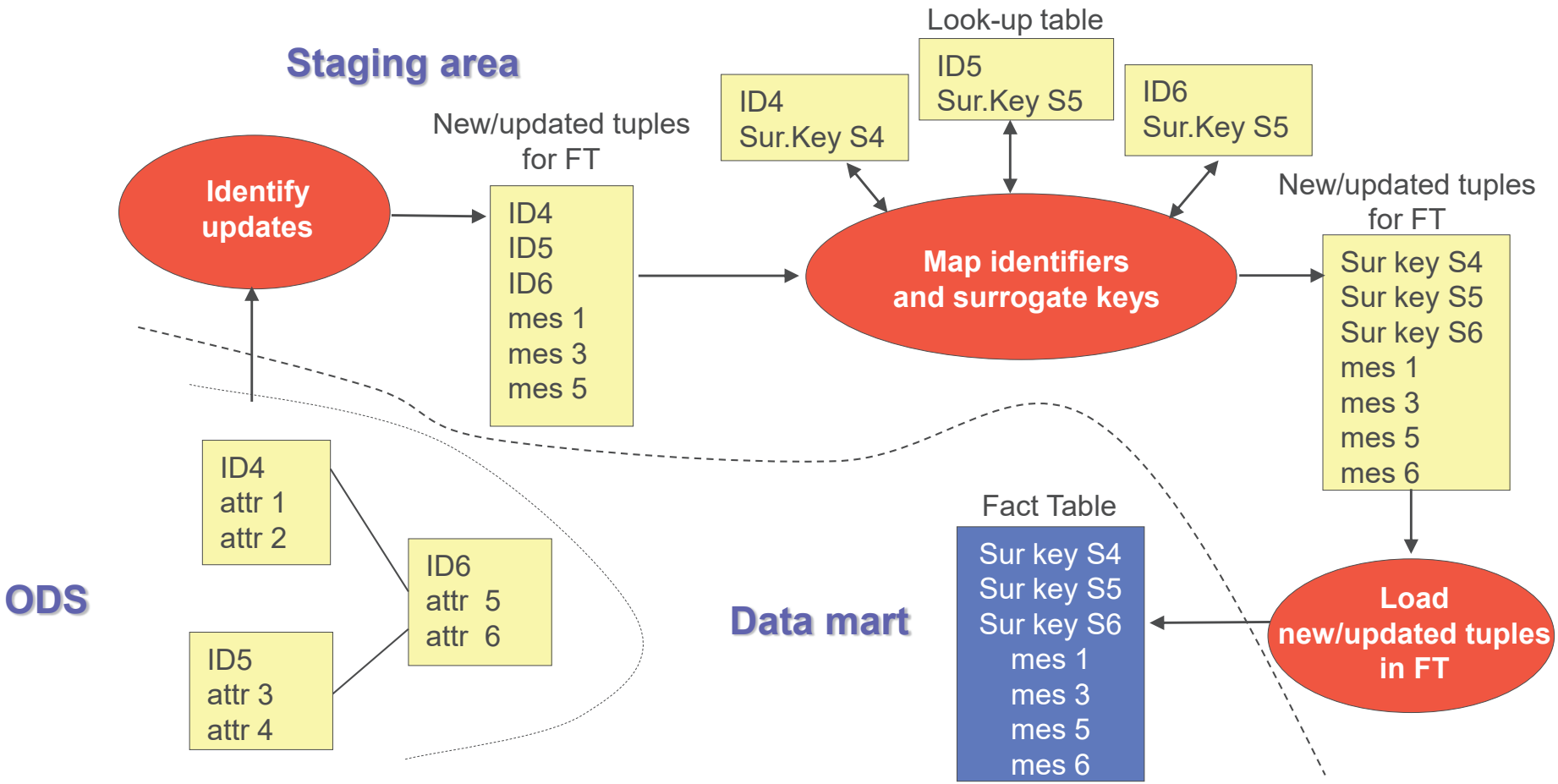
Dimension table loading



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

Elena Baralis
Politecnico di Torino

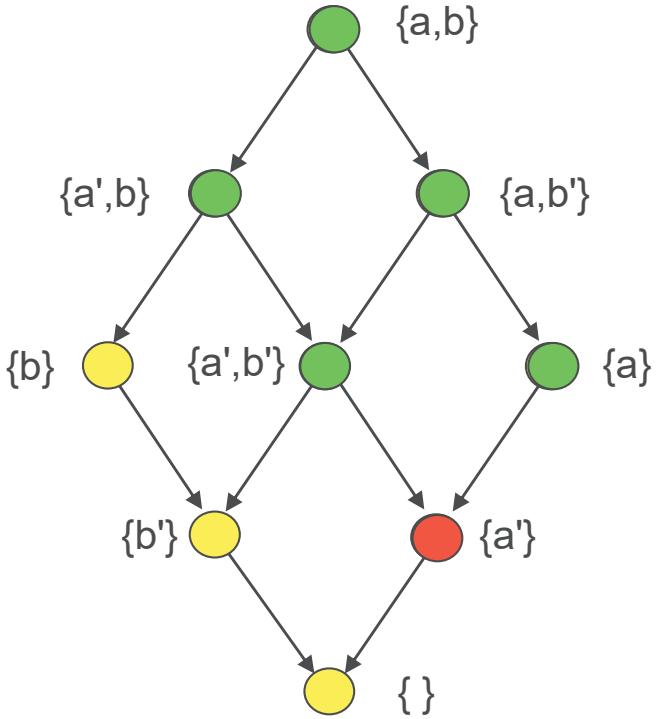
Fact table loading



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

Elena Baralis
Politecnico di Torino

Materialized view loading



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

Elena Baralis
Politecnico di Torino