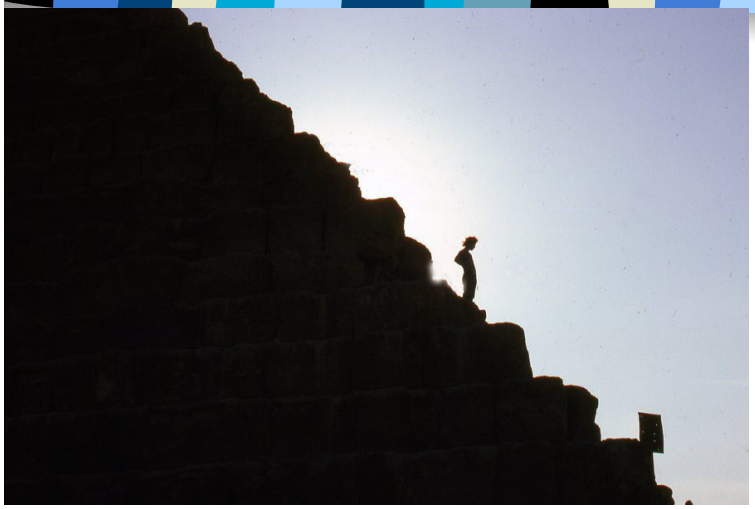
Data Warehouse and Beyond: Designing the BI Pyramid

Stefano Rizzi
DISI - University of Bologna - Italy
stefano.rizzi@unibo.it

Outline

- Climbing up the BI pyramid
- Data warehouse
  - ✓ Architectures
  - ✓ A design methodology
- What-if Analysis
  - ✓ Models and techniques
  - ✓ A design methodology
- Social business intelligence
  - ✓ Architectural framework
  - ✓ A design methodology

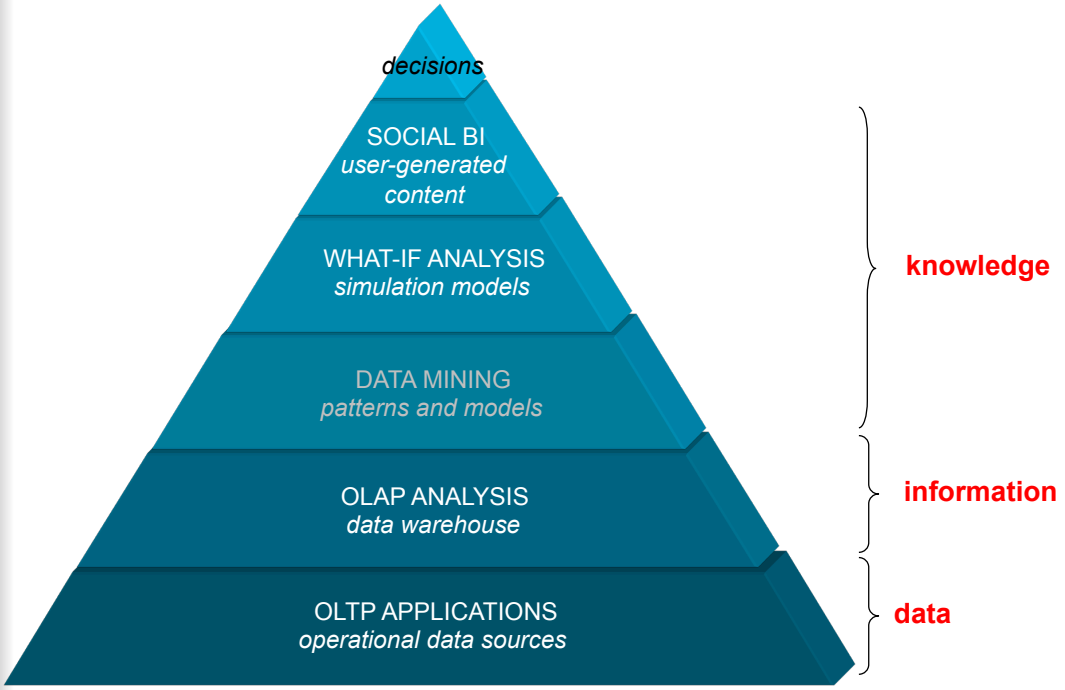
Climbing up the BI Pyramid



Business intelligence

- A set of tools and techniques that enable a company to transform its business data into timely and accurate information, so as to derive the knowledge necessary for the decisional process
  - ✓ Business intelligence systems are used by decision makers to get a comprehensive knowledge of the business and of the factors that affect it, as well as to define and support their business strategies
  - ✓ The goal is to enable data-based decisions aimed at gaining competitive advantage, improving operative performance, responding more quickly to changes, increasing profitability and, in general, creating added value for the company

The BI pyramid



Data Warehouse

The Data Warehouse

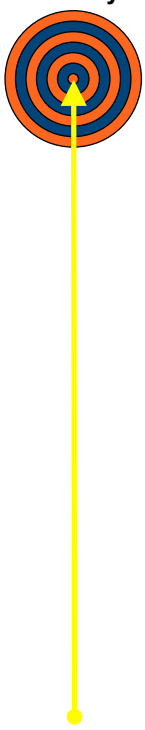
- *A data warehouse is a collection of information that supports decision-making processes*
  - ✓ *It is subject-oriented*
  - ✓ *It is integrated and consistent*
  - ✓ *It shows its evolution over time and it is not volatile*

Features of data warehouses

- accessibility to users not familiar with IT and data structures
- integration of data based on a standard enterprise model
- query flexibility to maximize the advantages obtained from the existing information
- information conciseness allowing for target-oriented and effective analyses
- multidimensional representation giving users an intuitive and manageable view of information
- correctness and completeness of integrated data

In Universities...

Cross-analyses



- Accounting
  - ✓ monitor financial flows
  - ✓ analyze incomes and expenses by budget item
  - ✓ ...
- Teaching
  - ✓ monitor student flows to assess the ability to attract and keep students
  - ✓ monitor the didactic load of teachers
  - ✓ ...
- HR
  - ✓ analyze employees by role, department, age
  - ✓ analyze teachers by scientific area and Faculty
  - ✓ monitor turnover
  - ✓ ...
- Research
  - ✓ analyze scientific productivity of teachers
  - ✓ analyze project fundings by department
  - ✓ ...

Architectural requirements

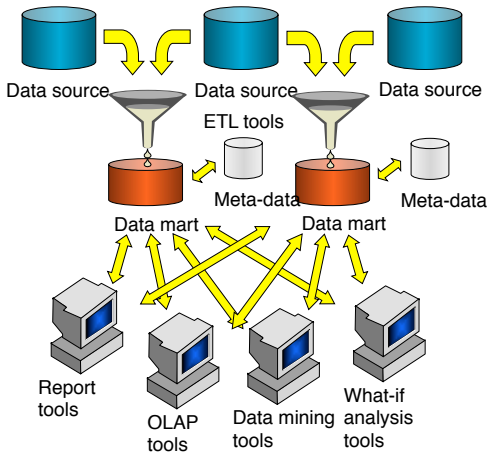
- ✓ **Separation** Analytical and transactional processing should be kept apart as much as possible
- ✓ **Scalability** Hardware and software architectures should be easy to upgrade as the data volume and the number of users progressively increase
- ✓ **Extendibility** The architecture should be able to host new applications and technologies without redesigning the whole system
- ✓ **Security** Monitoring accesses is essential because of the strategic data stored in data warehouses
- ✓ **Administerability** Data warehouse management should not be overly difficult

Architecture classification

- Independent data marts
- Data mart bus
- Hub-and-spoke

Independent data marts

- First approach to data warehousing
- Inconsistency issues

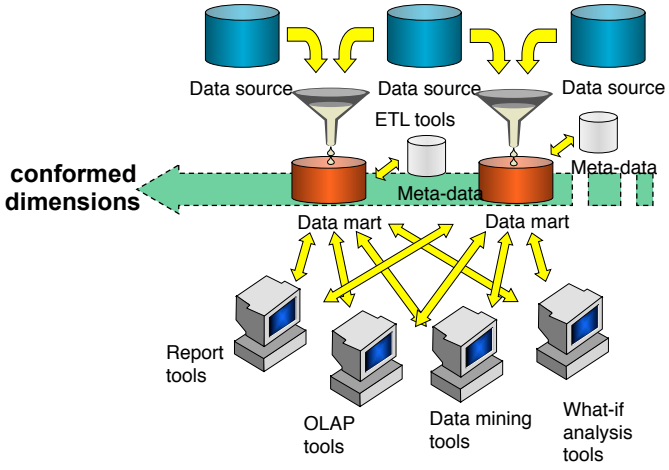


DATA MART:

A subset or an aggregation of the data stored to a primary data warehouse. It includes a set of information pieces relevant to a specific business area, corporate department, or category of users.

Data mart bus

- Approach suggested by Kimball
- Logical level integration
- “Enterprise view”



CONFORMED DIMENSIONS:

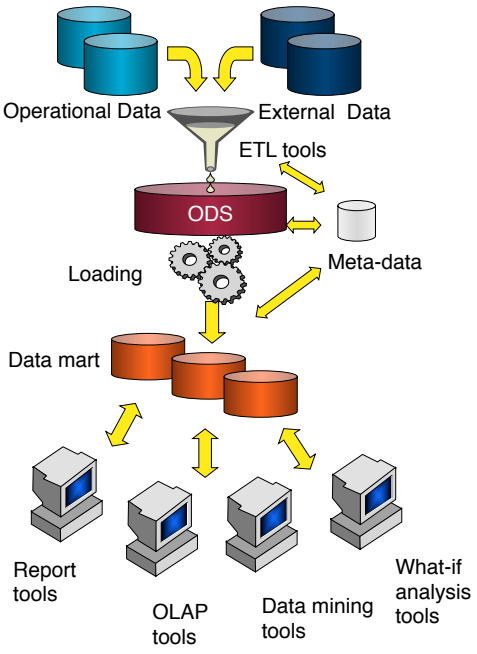
the main business dimensions shared by the whole enterprise, whose homogeneous design ensures the all data marts can be integrated

Hub-and-spoke

- One of the most used architectures in medium to large environments

OPERATONAL DATA STORE:

operational data obtained after integrating and cleansing source data. As a result, those data are integrated, consistent, appropriate, current, and detailed



Choosing an architecture

- Information interdependence among organizational units in company
  - ✓ encourages the adoption of enterprise-wide architectures
- Urgency of the data warehousing project
  - ✓ encourages the adoption of “fast” architectures
- Constraints on economic and human resources
- Role of the project within the business strategy
  - ✓ independent data marts vs. hub-and-spoke
- Compatibility with existing platforms
- Skills of the IT staff
- Organizational position of the sponsor of the project
  - ✓ enterprise architectures vs. departmental architectures

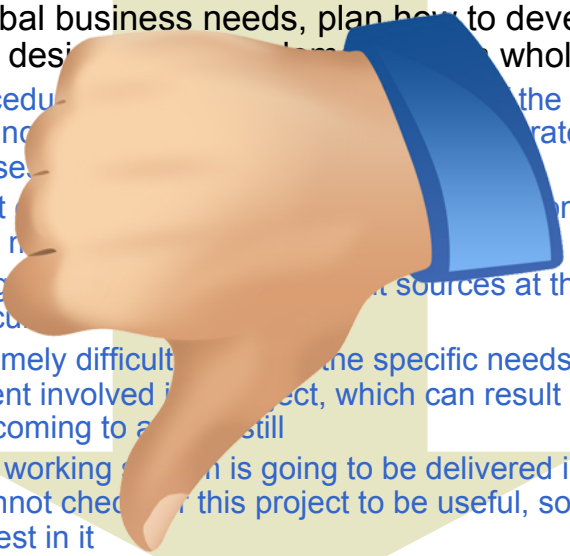
Data warehouse design

- Building a DW is a very complex task, which requires an accurate planning aimed at devising satisfactory answers to organizational and architectural questions
- The reports of DW project failures state that a major cause lies in the absence of a global view of the design process: in other terms, in the absence of a design methodology
- Methodologies are created by closely studying similar experiences and minimizing the risks for failure by basing new approaches on a constructive analysis of the mistakes made previously

M. Golfarelli, S. Rizzi.
***Data Warehouse Design: Modern Principles and Methodologies*.**
McGraw-Hill, 2009

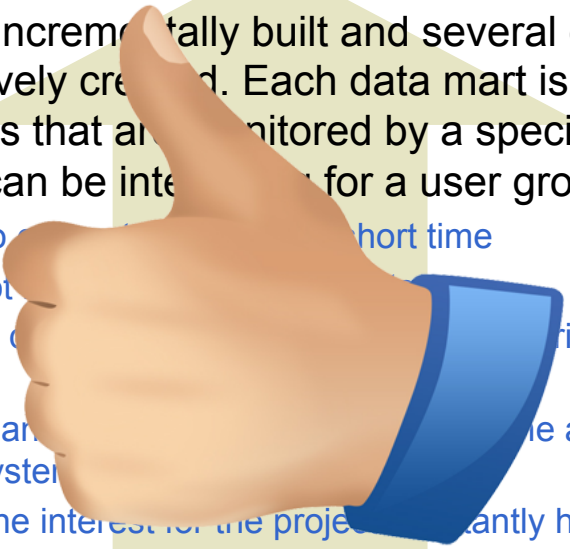
Top-down approach

- Analyze global business needs, plan how to develop a data warehouse, desi… whole
  - This procedu… the goal to achieve, and in princ… rated data warehouse…
  - High-cost … ns discourage company r… of projects
  - Analyzing … sources at the same time is a very difficu…
  - It is extremely difficult … he specific needs of every department involved i… ect, which can result in the analysis process coming to a… still
  - Since no working … is going to be delivered in the short term, users cannot chec… this project to be useful, so they lose trust and interest in it

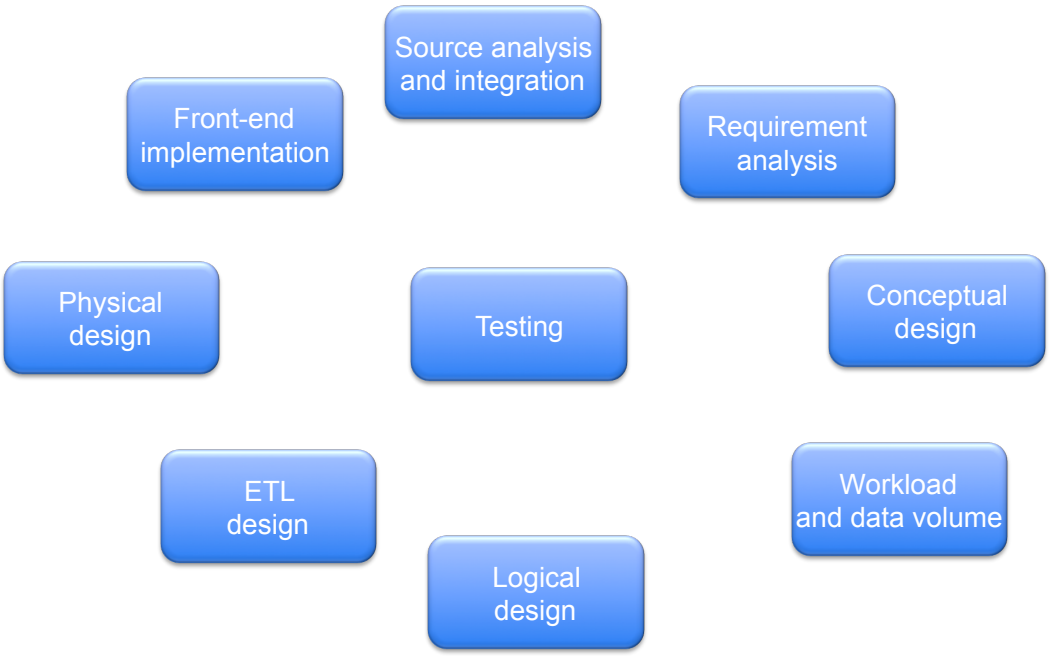


Bottom-up approach

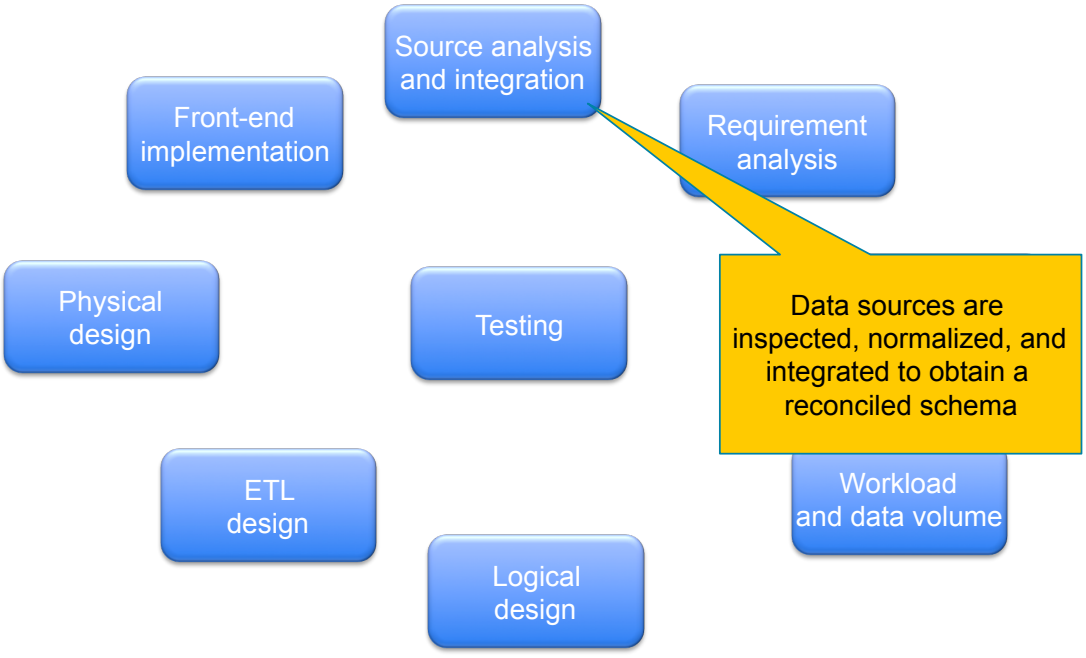
- DWs are increme… tally built and several data marts are iteratively cre… d. Each data mart is based on a set of facts that ar… nitored by a specific division and that can be inte… for a user group
  - Leads to … hort time
  - Does not …
  - Enables … rise area at a time
  - Gives mar… e actual benefits of the syste…
  - Keeps the inter… the proje… antly high
  - May determine a partial vision of the business domain



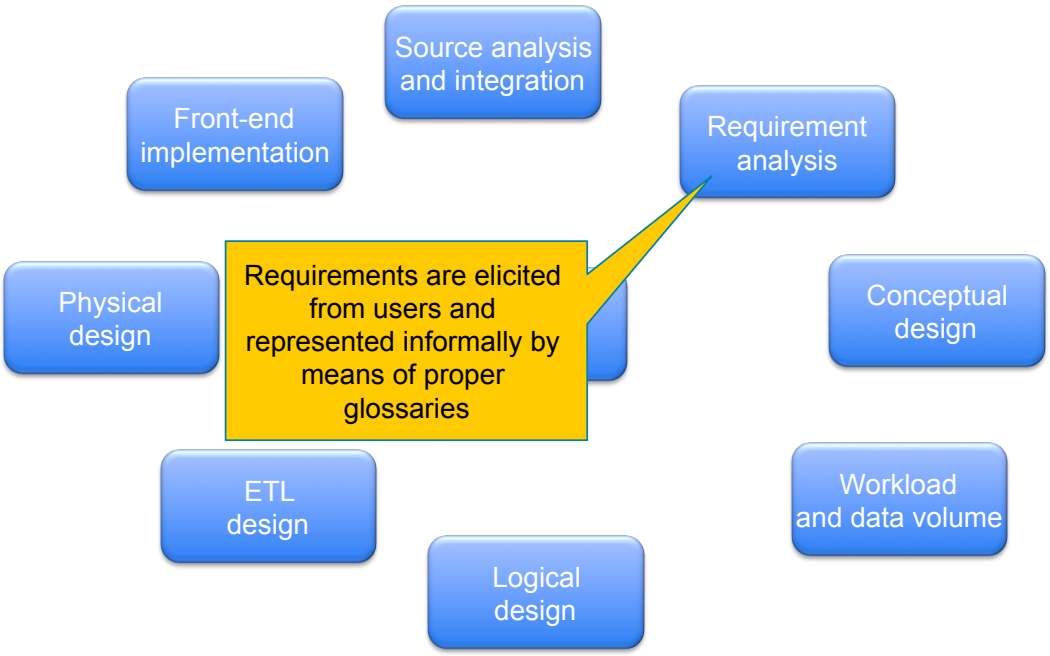
Data mart design phases



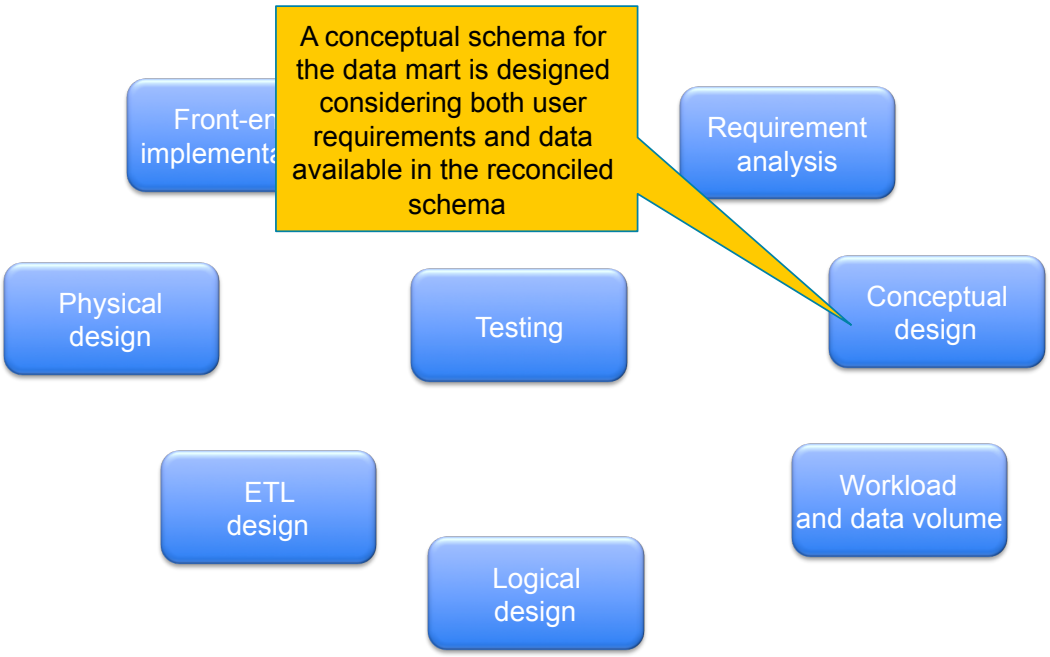
Data mart design phases



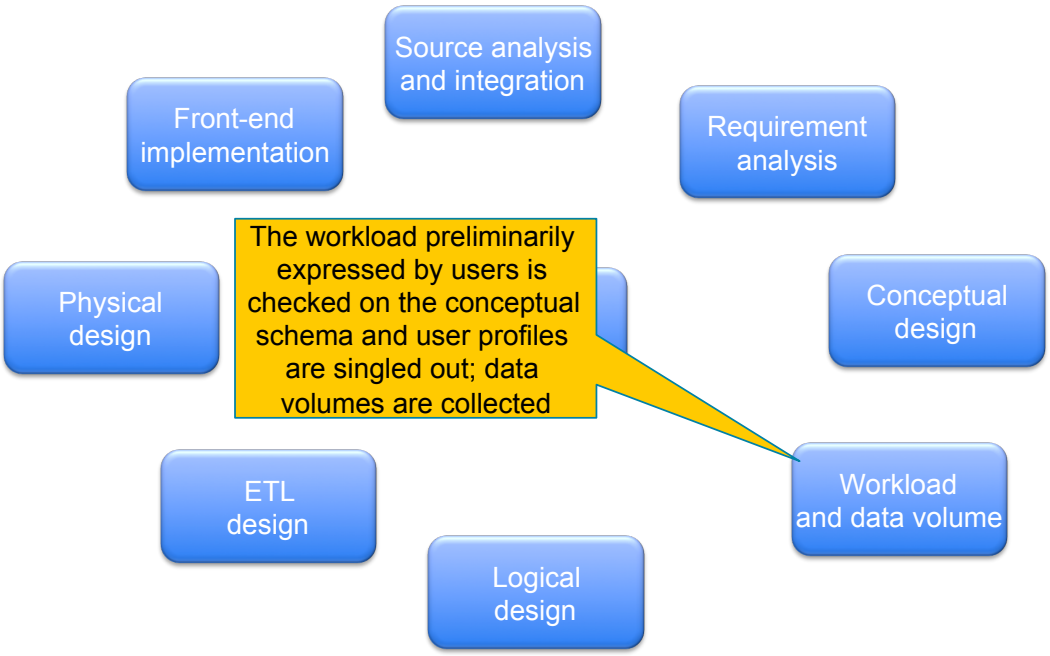
Data mart design phases



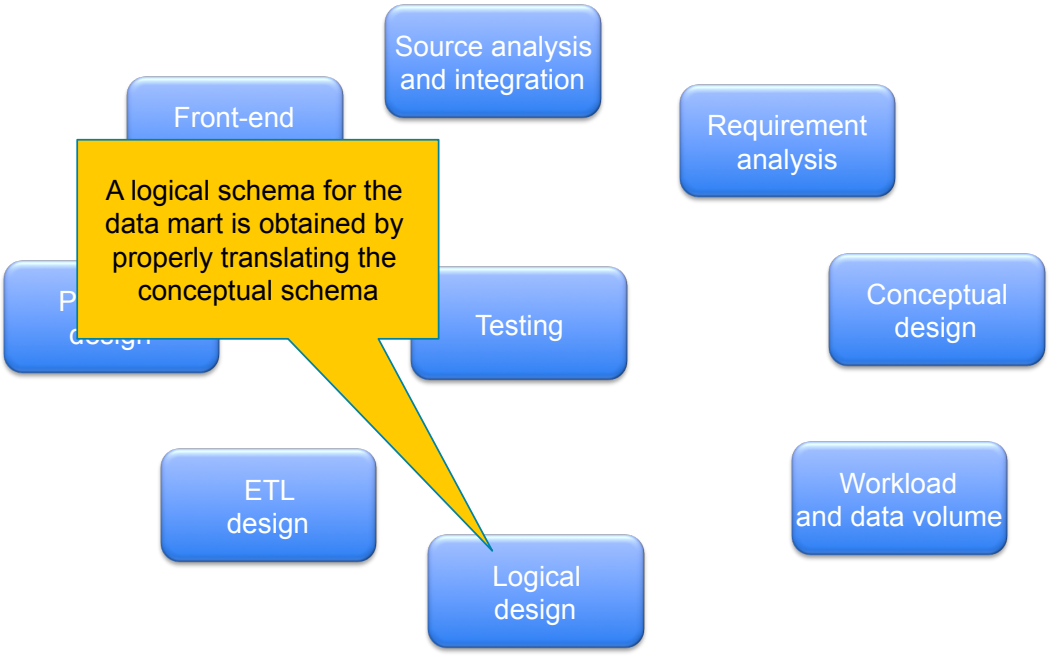
Data mart design phases



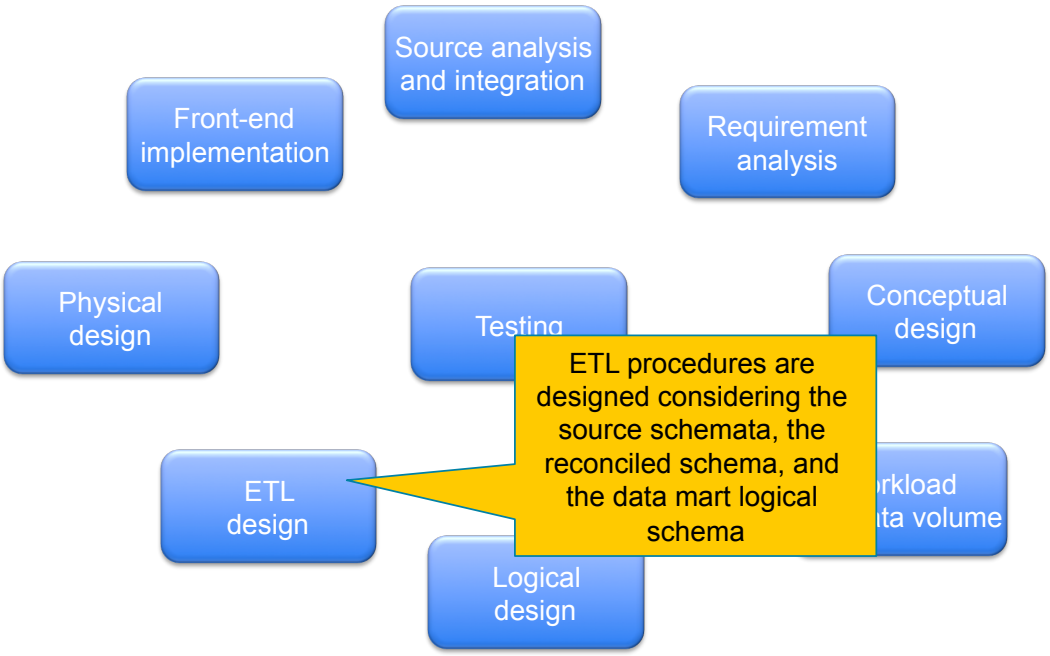
Data mart design phases



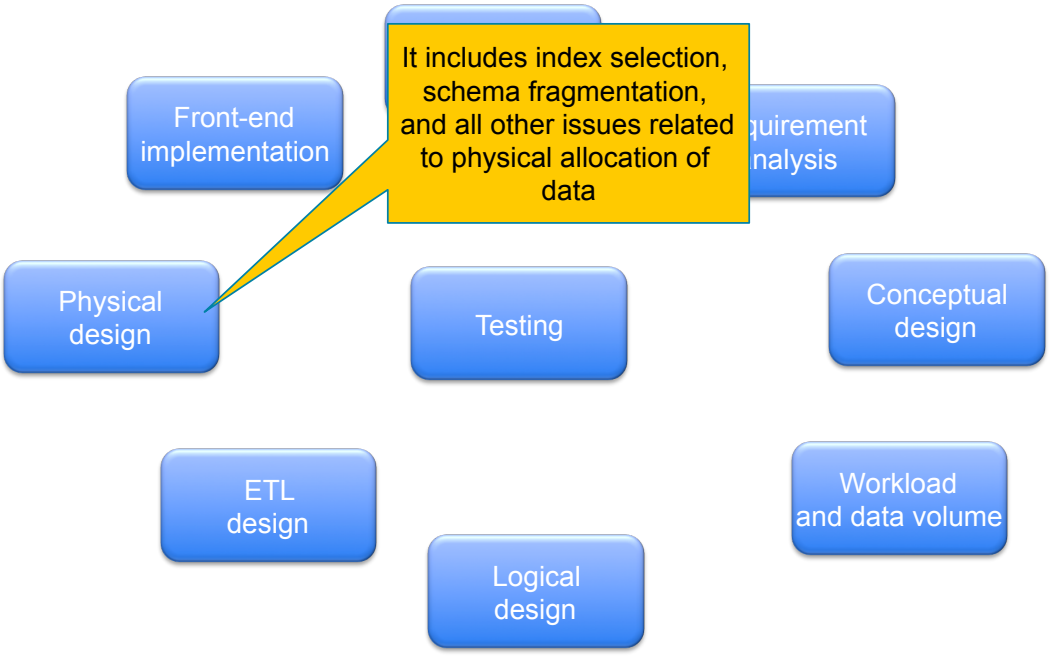
Data mart design phases



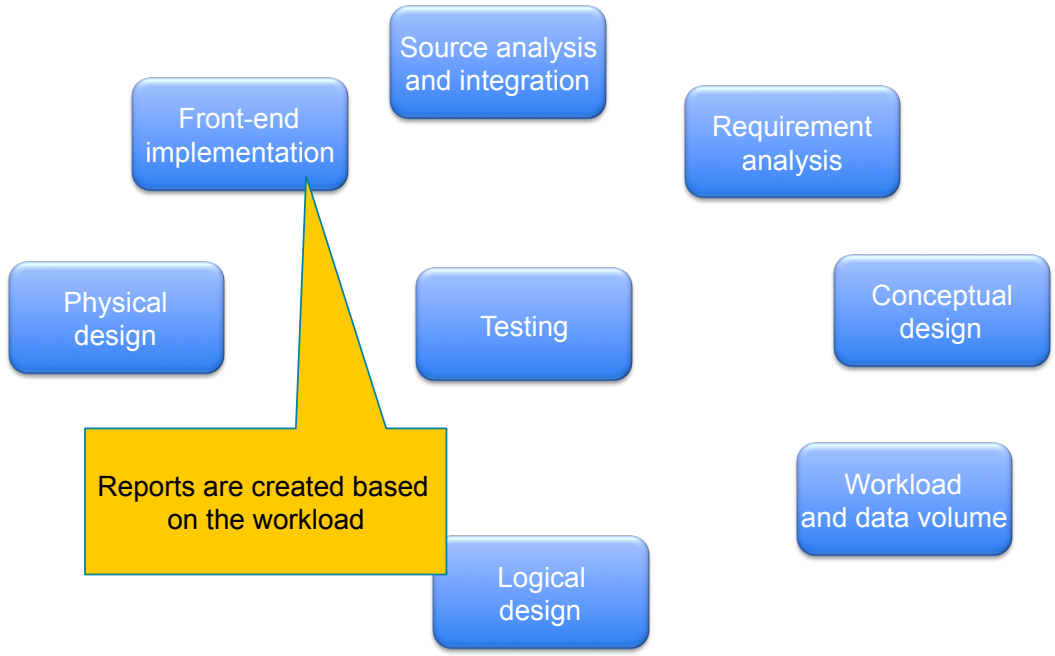
Data mart design phases



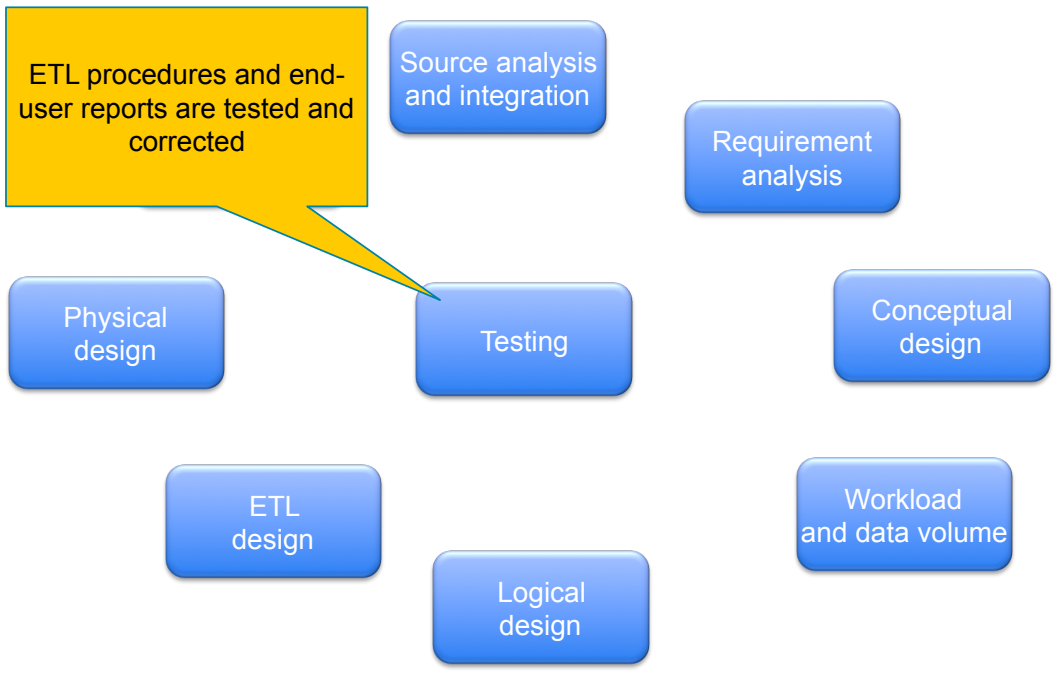
Data mart design phases



Data mart design phases



Data mart design phases



Methodological scenarios

- *Supply-driven approach*
  - ✓ data marts are designed based on a close operational data source analysis
  - ✓ user requirements show designers which groups of data, relevant for decision-making processes, should be selected
- *Demand-driven approach*
  - ✓ it begins with the definition of information requirements of data mart users
  - ✓ the problem of how to map those requirements onto existing data sources is addressed at a later stage, when ETL procedures are implemented

Which formalism for conceptual design?

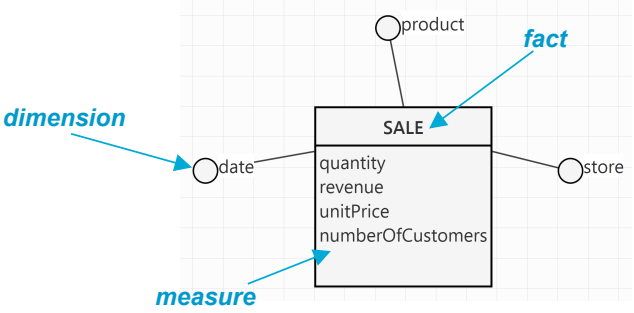
- While it is now universally recognized that a data mart is based on a multidimensional view of data, there is still no agreement on how to implement its conceptual design
- Use of the Entity-Relationship model is quite widespread throughout companies as a conceptual tool for standard documentation and design of relational databases, but *it cannot be used to model DWs*
- In some cases, designers base their data marts design on the logical level—that is, they directly define star schemata that are the standard ROLAP implementation of the multidimensional model. But a star schema is nothing but a relational schema; *it contains only the definition of a set of relations and integrity constraints!*

The Dimensional Fact Model

- The DFM is a graphical conceptual model for data mart design, devised to:
  1. lend effective support to conceptual design
  2. create an environment in which user queries may be formulated intuitively
  3. make communication possible between designers and end users with the goal of formalizing requirement specifications
  4. enable early testing of requirements
  5. build a stable platform for logical design (*independently of the target logical model*)
  6. provide clear and expressive design documentation
- The conceptual representation generated by the DFM consists of a set of *fact schemata* that basically model facts, measures, dimensions, and hierarchies

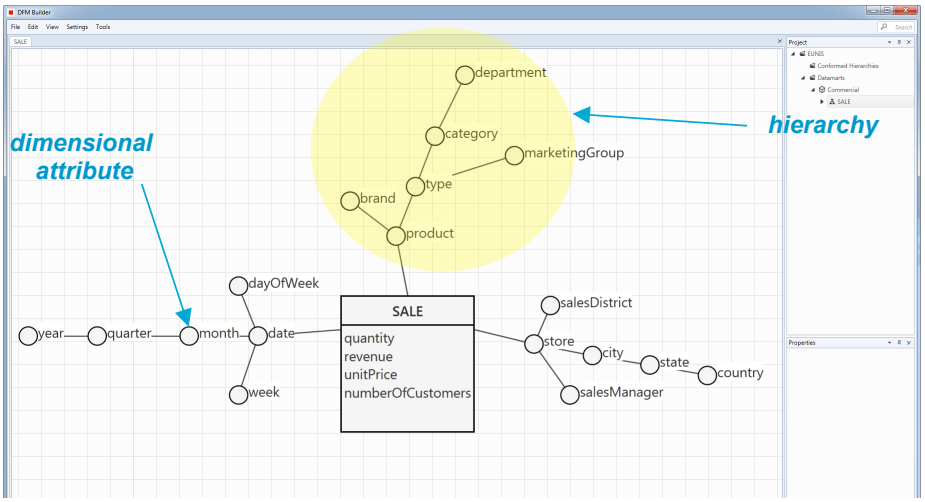
DFM: basic concepts

- A *fact* is a concept relevant to decision-making processes. It typically models a set of events taking place within a company. It is essential that a fact have dynamic properties or evolve in some way over time
- A *measure* is a numerical property of a fact and describes a quantitative fact aspect that is relevant to analysis
- A *dimension* is a fact property with a finite domain and describes an analysis coordinate of the fact.

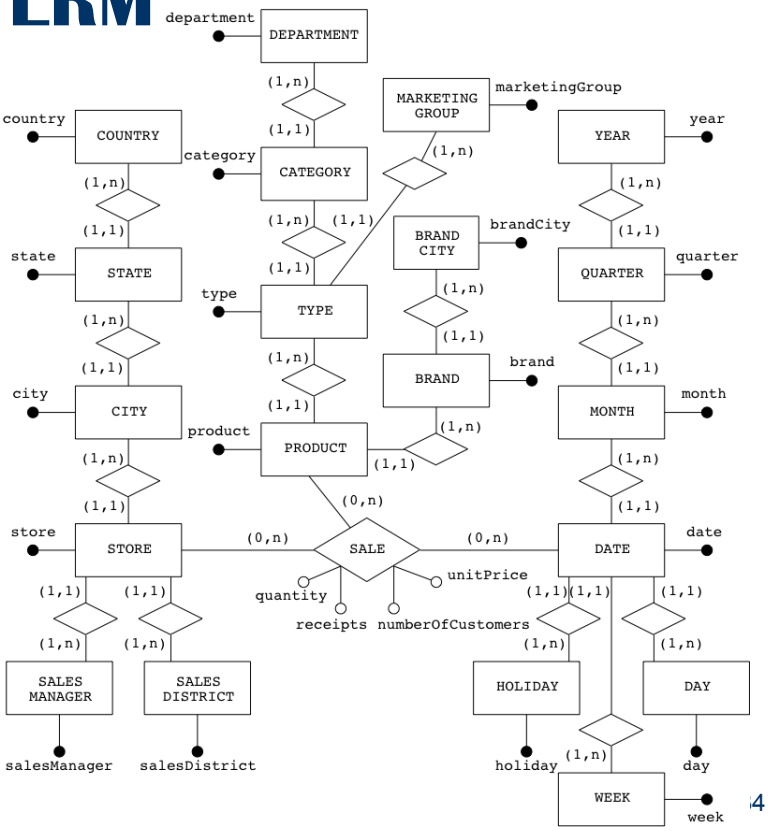


DFM: basic concepts

- The general term *dimensional attributes* stands for the dimensions and other possible attributes, always with discrete values, that describe them
- A *hierarchy* is a directed tree whose nodes are dimensional attributes and whose arcs model many-to-one associations between dimensional attribute pairs

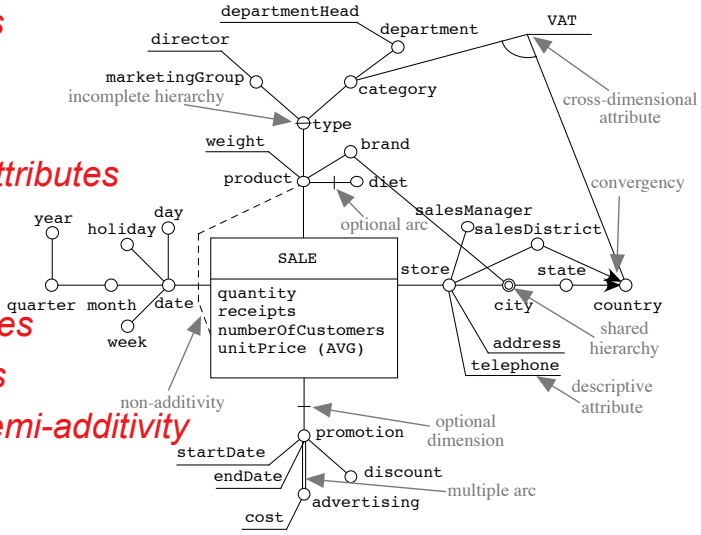


DFM vs. ERM



DFM: advanced concepts

- The DFM also supports
  - ✓ *descriptive attributes*
  - ✓ *optional arcs*
  - ✓ *convergences*
  - ✓ *cross-dimensional attributes*
  - ✓ *shared hierarchies*
  - ✓ *multiple arcs*
  - ✓ *incomplete hierarchies*
  - ✓ *recursive hierarchies*
  - ✓ *non-additivity and semi-additivity*



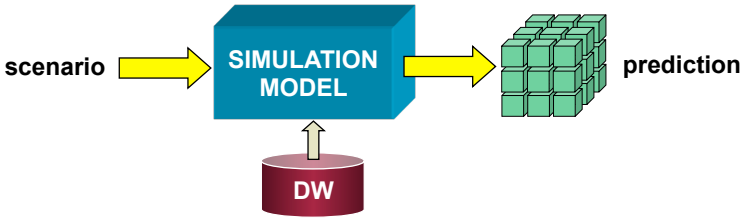
What-if Analysis

What-if analysis

- DWs support analyses of past data, but give no view of future trends
- Decision makers need to evaluate beforehand the impact of a strategic or tactical move
  - ✓ "How would my profits change if I ran a 3×2 promotion for one week on some product on sale?"
    - Modeling the behavior of the customers
    - Modeling the side effects on similar product sales in the same week (*cannibalization*)
    - Modeling the side effects on the product sales in the next weeks

What-if analysis

- What-if analysis is a data-intensive simulation whose goal is to inspect the behavior of a complex system under some given hypotheses (called *scenarios*)
- What-if analysis measures how the changes in a set of independent variables affect the values of a set of dependent variables with reference to a simulation model; this model gives a simplified representation of business, tuned on historical enterprise data

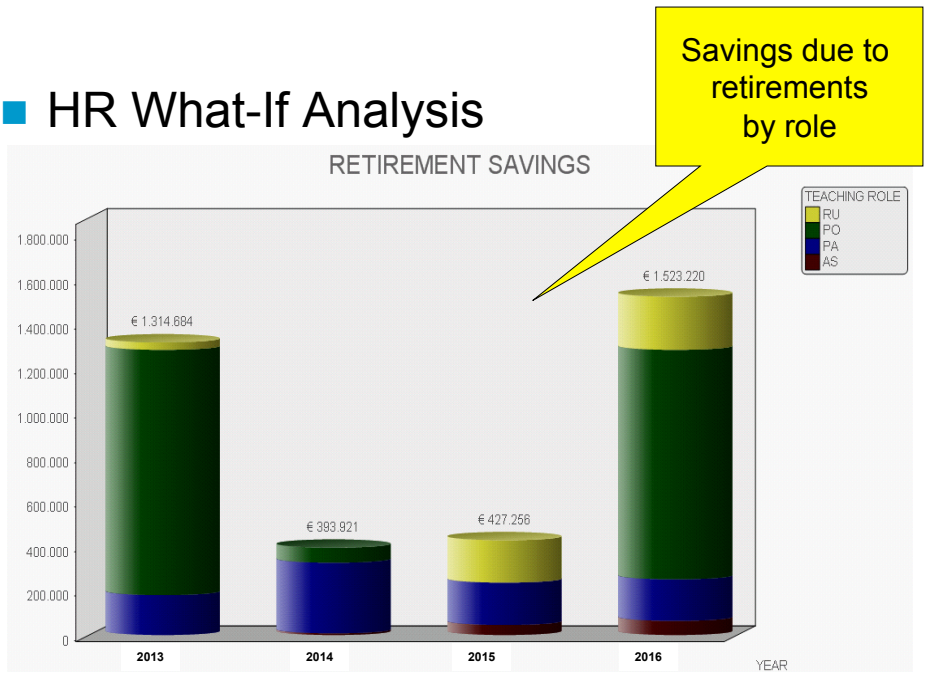


Expressing vs. building the simulation model

- Techniques to *express* the simulation model
  - E.g.: equations, rules, algorithms, correlation matrices, ...
- Techniques to *build* the simulation model
  - ✓ Statistical techniques: they derive a model starting from the behaviour of the system in the past
    - E.g.: regression, data mining
    - they do not capture the causes of phenomena, only their effects
    - they may fail on a complex system if historical data do not comprehensively describe the system behaviour
  - ✓ Judgment techniques: they analyze and formalize the cause-effect relationships that rule the system behaviour
    - E.g.: joint analysis and role-playing game
    - they produce more general and accurate models
    - they can hardly be applied to complex systems

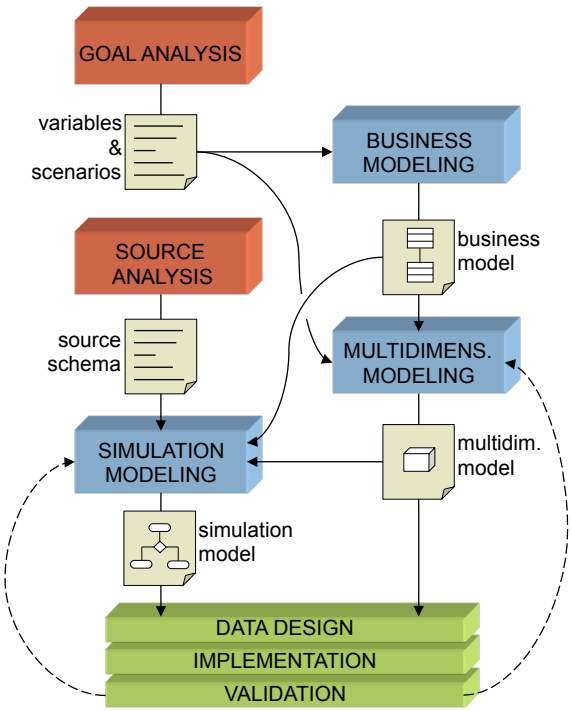
In Universities...

- HR What-If Analysis



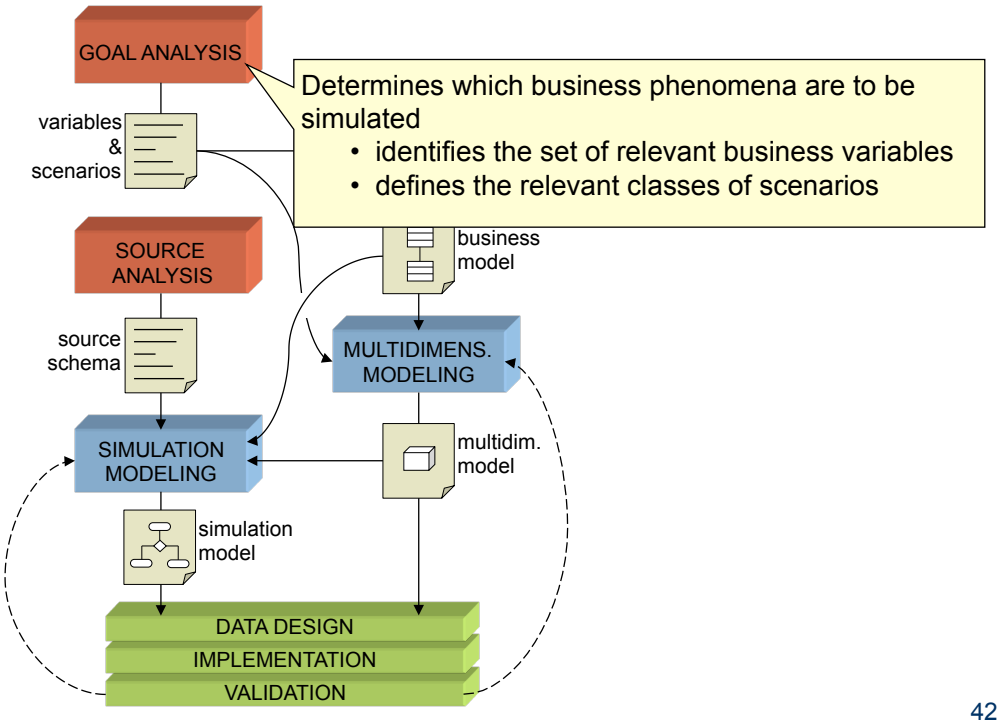
- Tuition Fees Analysis ….

Methodological sketch

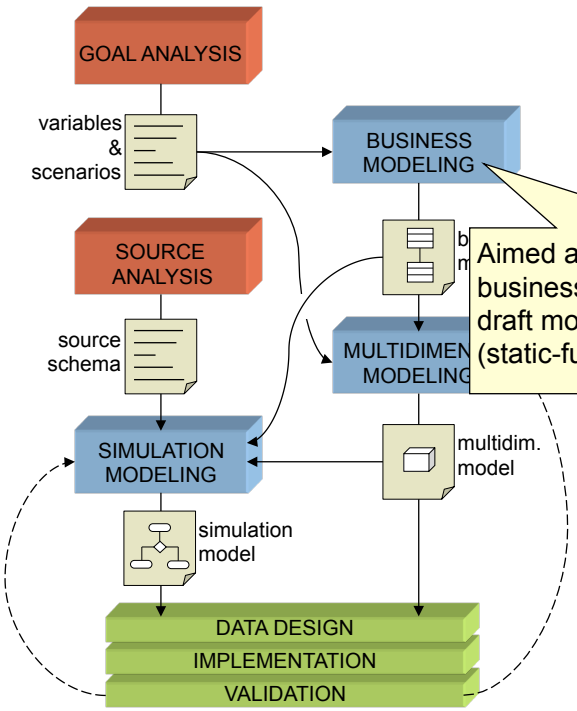


M. Golfarelli, S. Rizzi. What-if simulation modeling in business intelligence. International Journal of Data Warehousing and Mining, vol. 5, n. 4, pp. 24-43, 2009

Methodological sketch

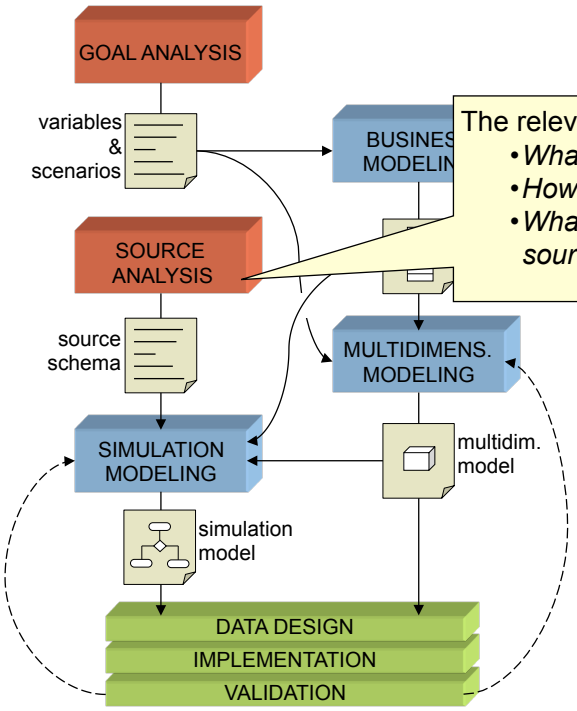


Methodological sketch



Aimed at better understanding the business phenomenon by building a draft model of the application domain (static-functional-dynamic)

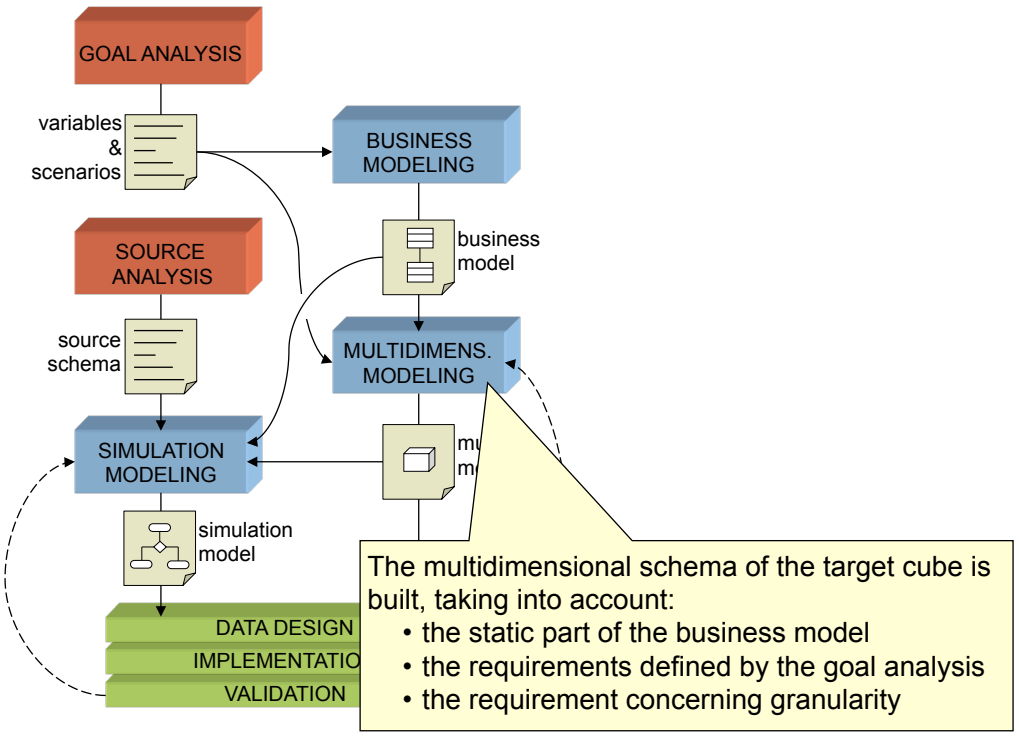
Methodological sketch



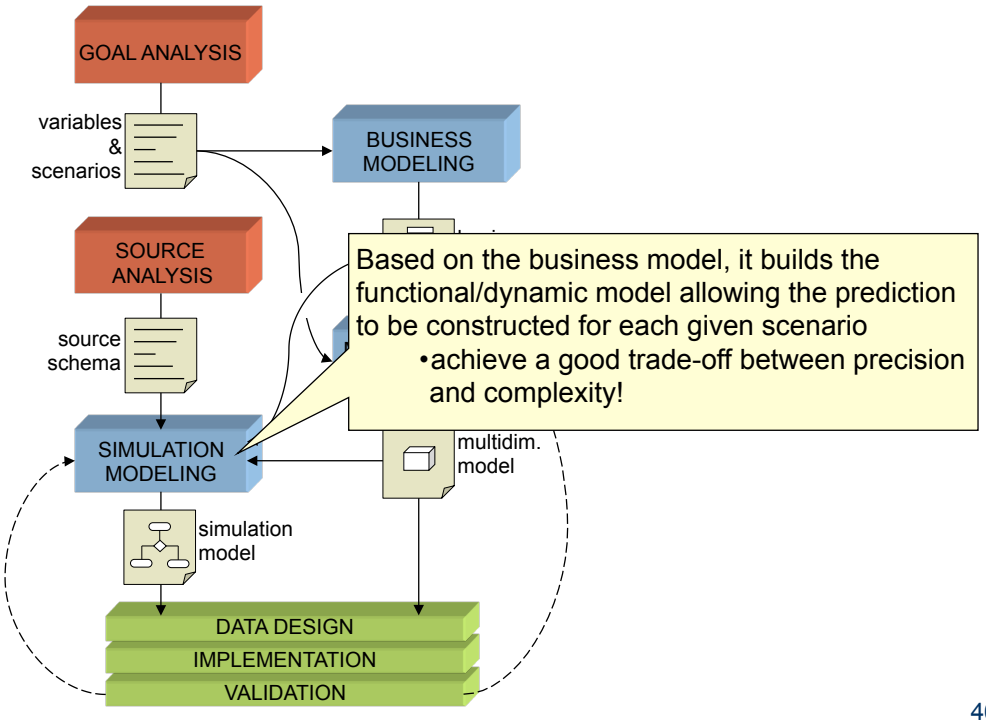
The relevant data sources are analyzed

- *What information is available?*
- *How is it structured?*
- *What is the quality of each data source?*

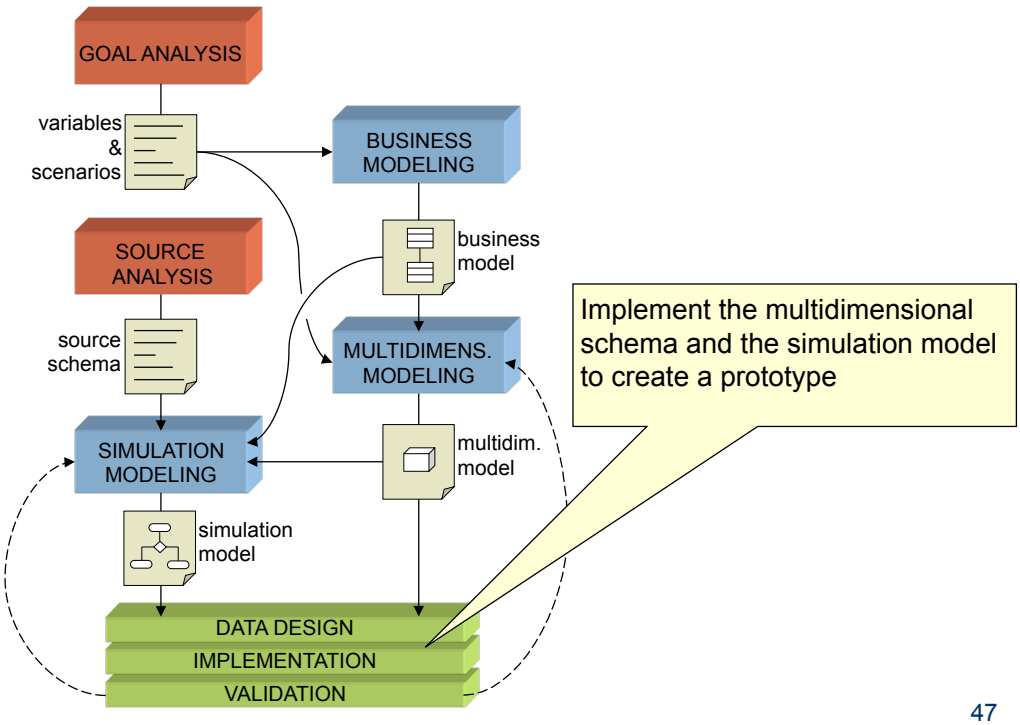
Methodological sketch



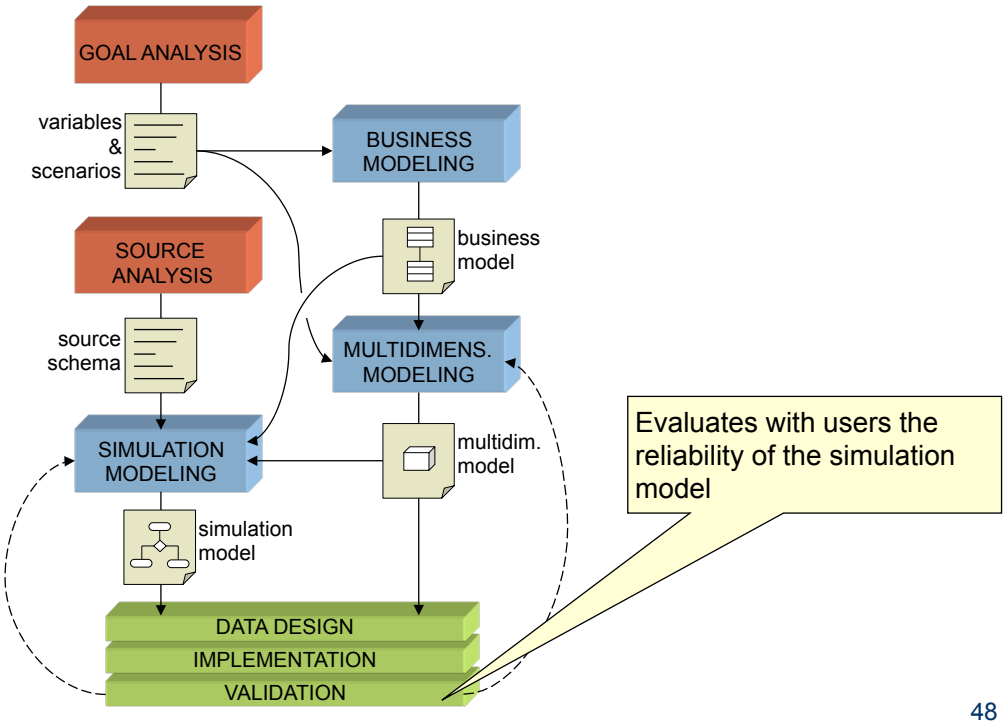
Methodological sketch



Methodological sketch



Methodological sketch



Social BI

Motivation

- Social networks and portable devices enabled simplified and ubiquitous forms of communication which contributed, during the last decade, to a boost in the voluntary sharing of personal information
- As a result, an enormous amount of user-generated content related to people's tastes, thoughts, and actions has been made available in the form of preferences, opinions, geolocation, etc.
- This huge wealth of information is raising an increasing interest from decision makers because it can give them a timely perception of the market mood and help them explain the phenomena of business and society

A definition

- *Social BI is the emerging discipline that aims at effectively and efficiently combining corporate data with UGC to let decision-makers analyze and improve their business based on the trends and moods perceived from the environment*
- As in traditional BI, the goal of SBI is to enable powerful and flexible analyses for decision makers with a limited expertise in databases and ICT

An architecture for Social BI

E. Gallinucci, M. Golfarelli, S. Rizzi. Meta-stars: multidimensional modeling for social business intelligence. Proceedings 16th International Workshop on Data Warehousing and OLAP (DOLAP 2013), S. Francisco, USA, pp. 11-18, 2013

Sentiment analysis

Capability of determining

- ✓ the attitude of an *opinion holder* about a given topic
- ✓ the *polarity* or *bias* of a document or a single sentence

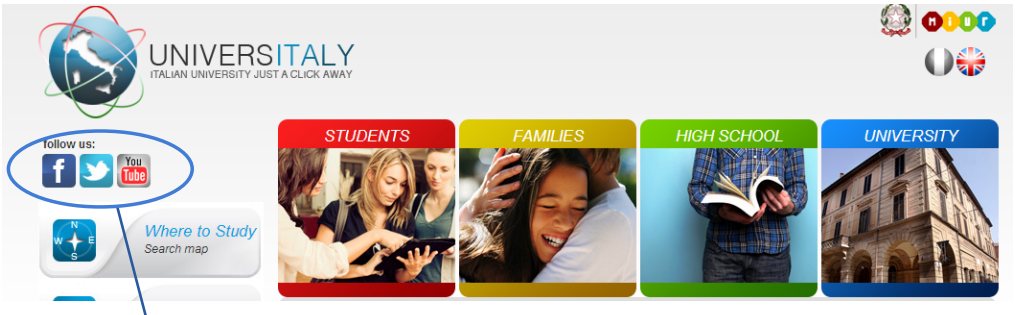
through

- ✓ automated identification
- ✓ extraction
- ✓ processing
- ✓ evaluation

of subjective information in the source document

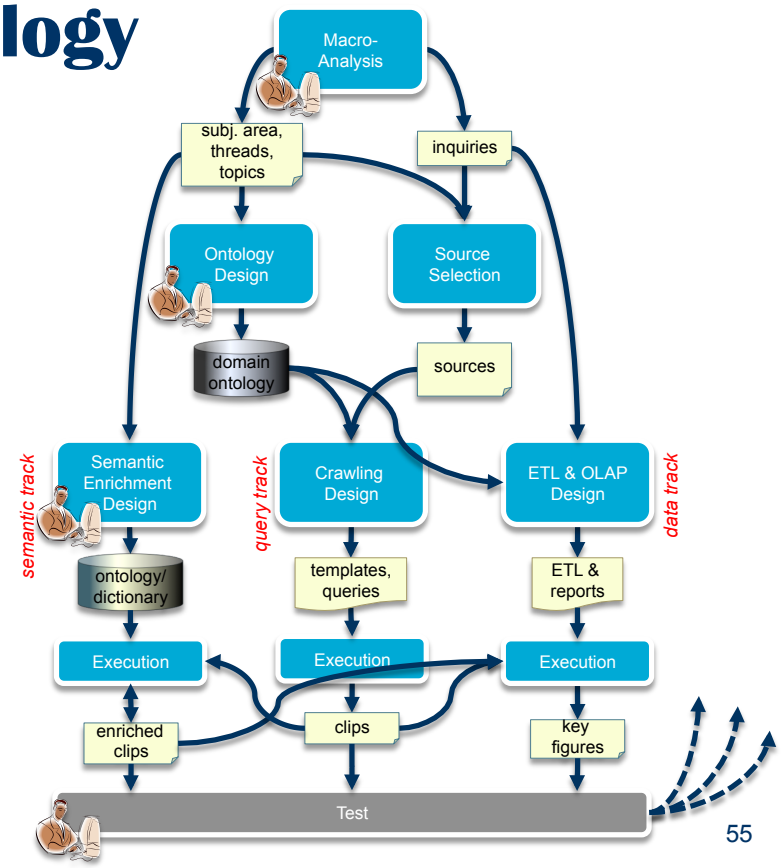
In Universities...

- Reputation of Universities



University Portal to promote Italian Higher Education

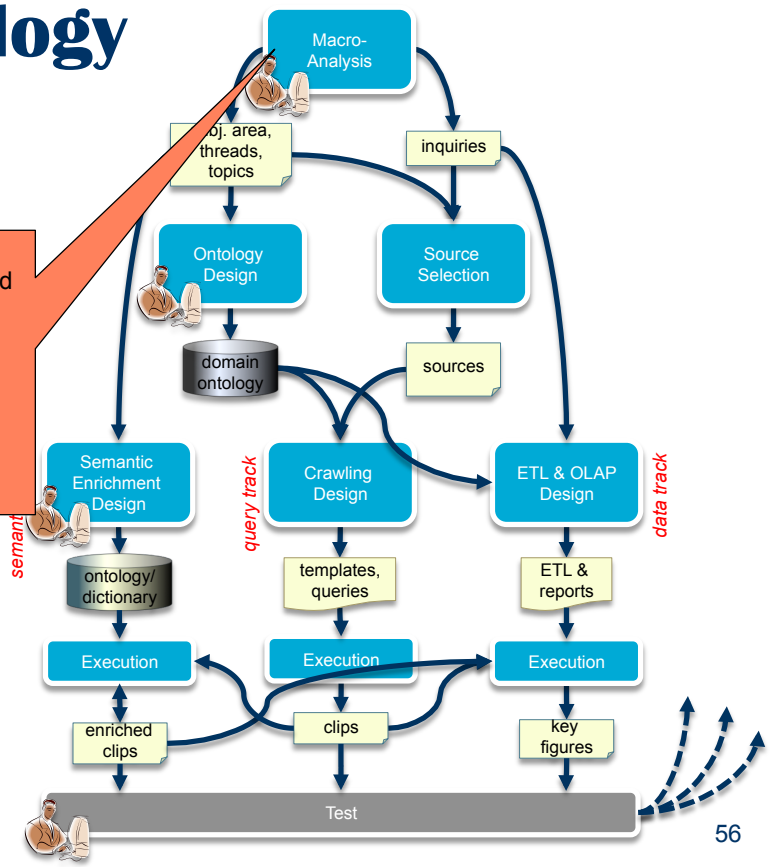
Methodology



Methodology

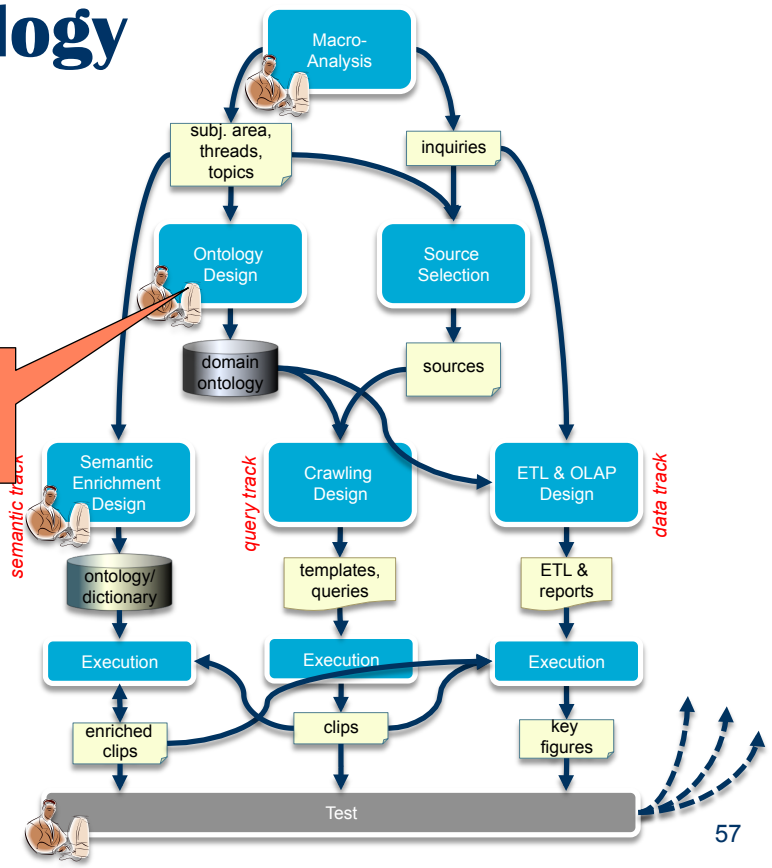
Users are interviewed to define the project scope and the set of inquiries the system will answer to

- An inquiry captures an informative need of a user; it is specified by *what*, *how*, and *where*



Methodology

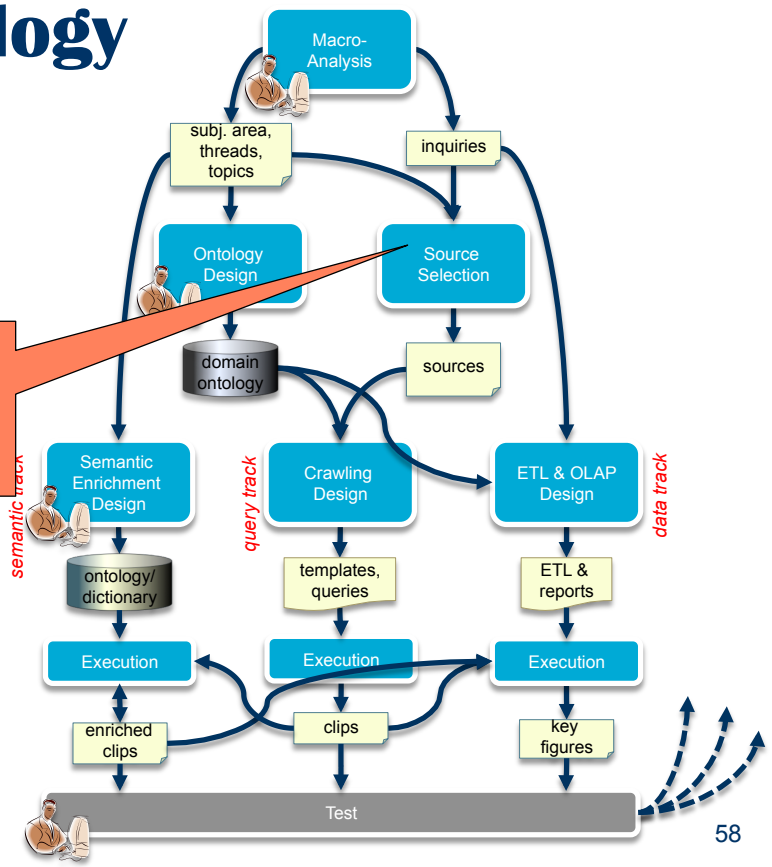
Customers work on themes and topics to build and refine the domain ontology that models the subject area



Methodology

It is aimed at identifying as many web domains as possible for crawling:

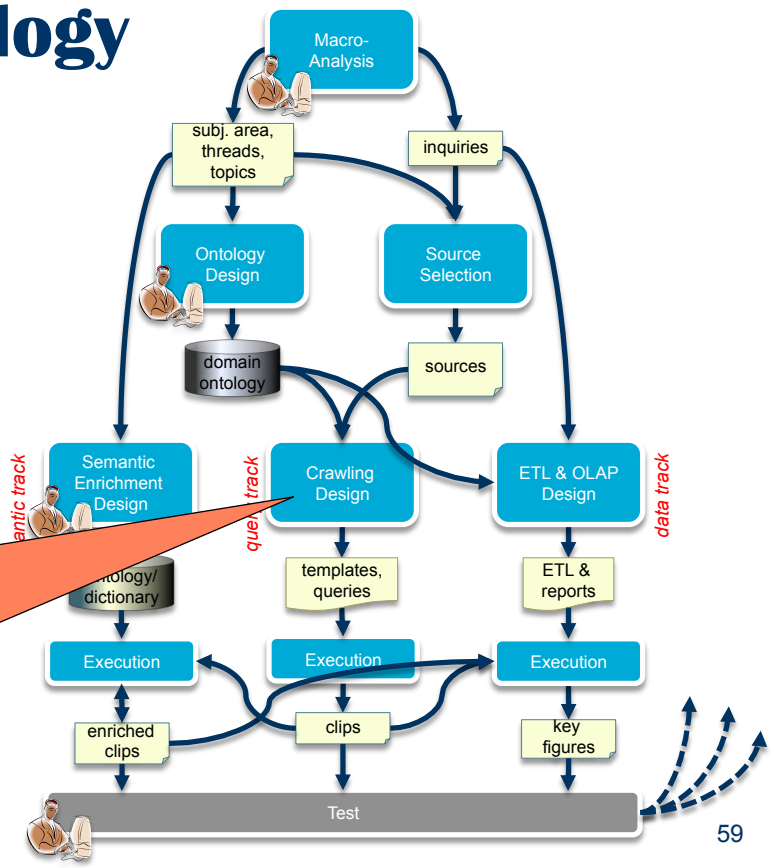
- *primary sources*
- *minor sources*



Methodology

Crawling design aims at retrieving in-topic clips by filtering off-topic clips out. A set of queries are created to search for relevant clips across the selected sources

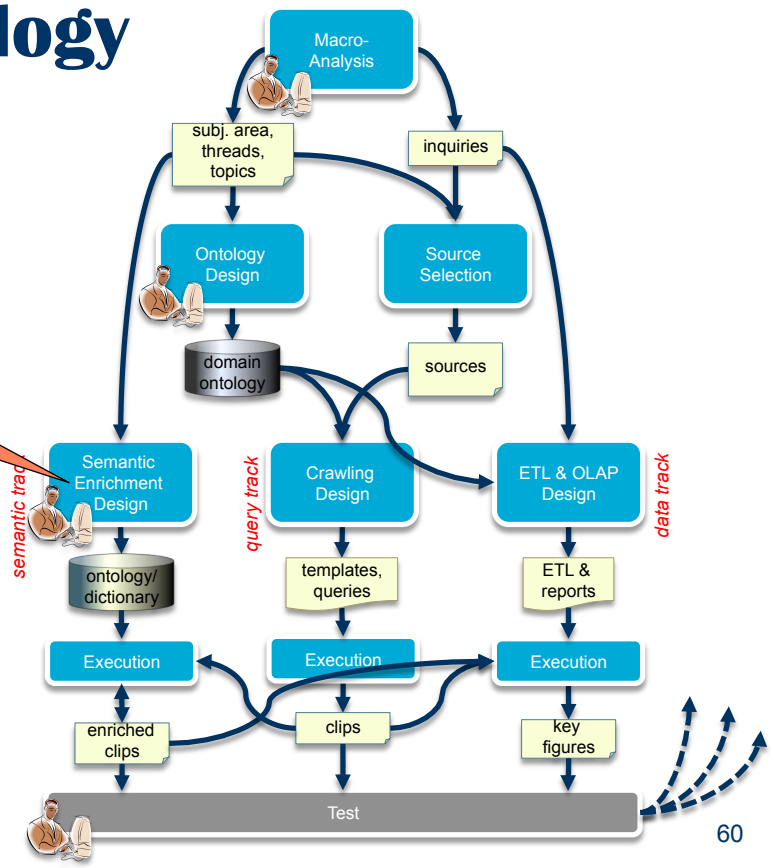
1. *Template design*
2. *Query design*
3. *Content relevance analysis*



Methodology

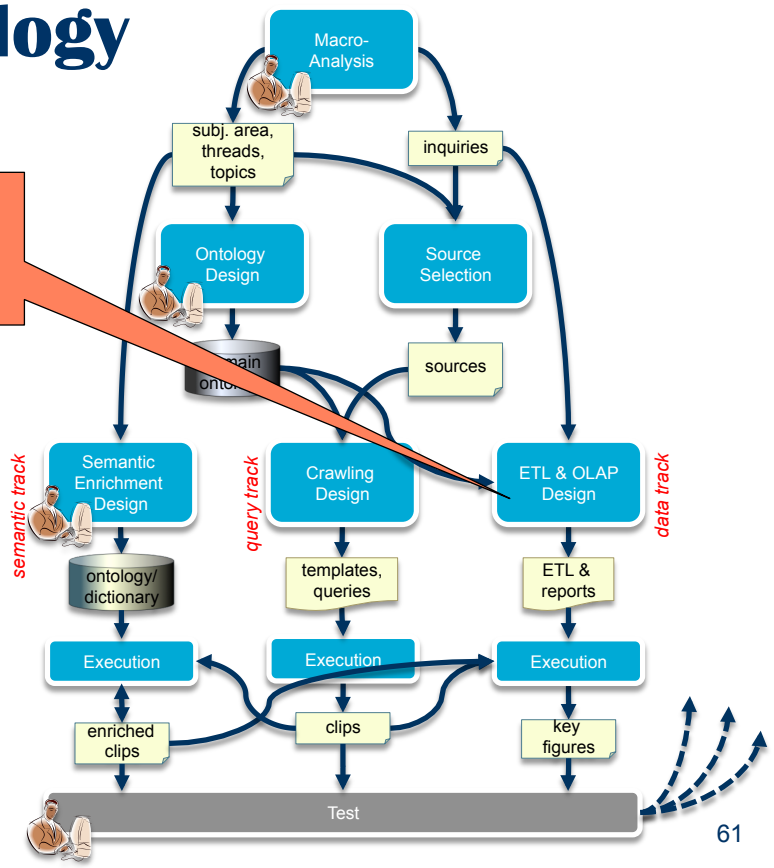
Increase the accuracy of text analytics so as to maximize the process effectiveness in terms of extracted entities and sentiment assigned to clips

- *Dictionary enrichment*
- *Inter-word relation definition*



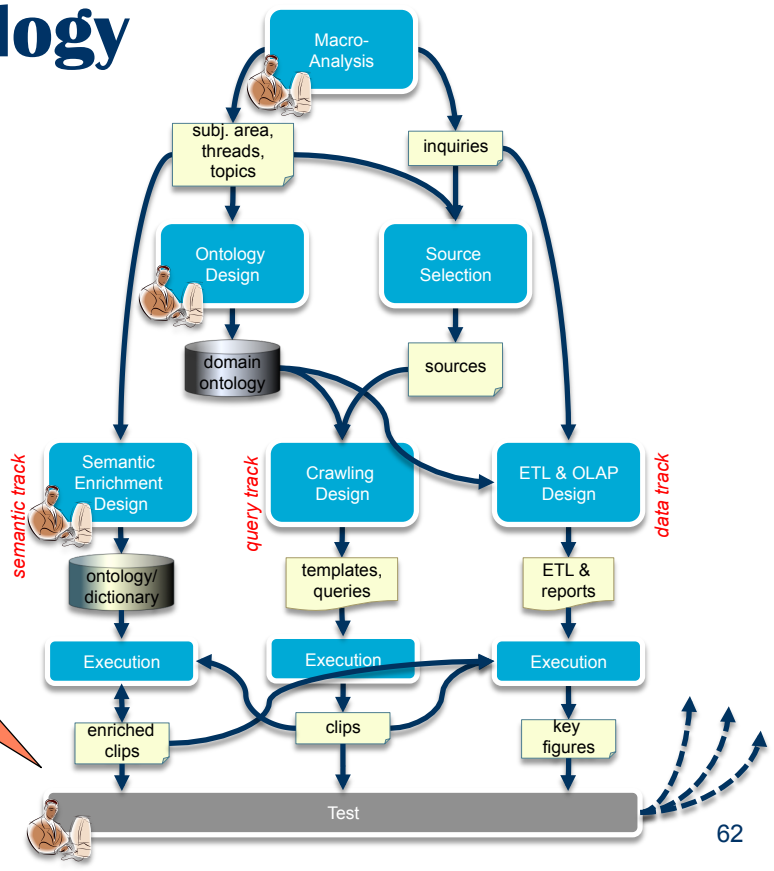
Methodology

1. ETL design and implementation
2. KPI design
3. Dashboard design



Methodology

Crawling queries are executed, the resulting clips are processed, and the reports are launched over the enriched clips



Conclusions: DW

- Adopting a structured design methodology based on conceptual design and on early testing ensures:
  - ✓ shorter design and validation times
  - ✓ better compliance with user requirements
  - ✓ availability of good-quality documentation
  - ✓ reduction of maintenance and evolution costs

Conclusions: what-if

- The diffusion of what-if analysis projects is surprisingly low
- Two main factors contribute to this:
  - ✓ Immature technology
    - The new generation of analytic tools are now compensating the technological gap
  - ✓ Design complexity
    - Complexity can be overcome by relying on pre-configured models (e.g., SAP-BPS is based on the business models captured by its ERP)

Conclusions: SBI

- Responsiveness in an SBI project is not a choice but rather a necessity, since the frequency of changes requires a tight involvement of domain experts to detect these changes and rapid iterations to keep the process well-tuned
- Such a frantic setting imposes a radical change in the project management approach with reference to traditional BI projects and a large effort to both end-users and developers
  - ✓ To reduce such effort, customers often outsource the activities yielding the worst trade-off between effort and added value for the SBI process

Questions?