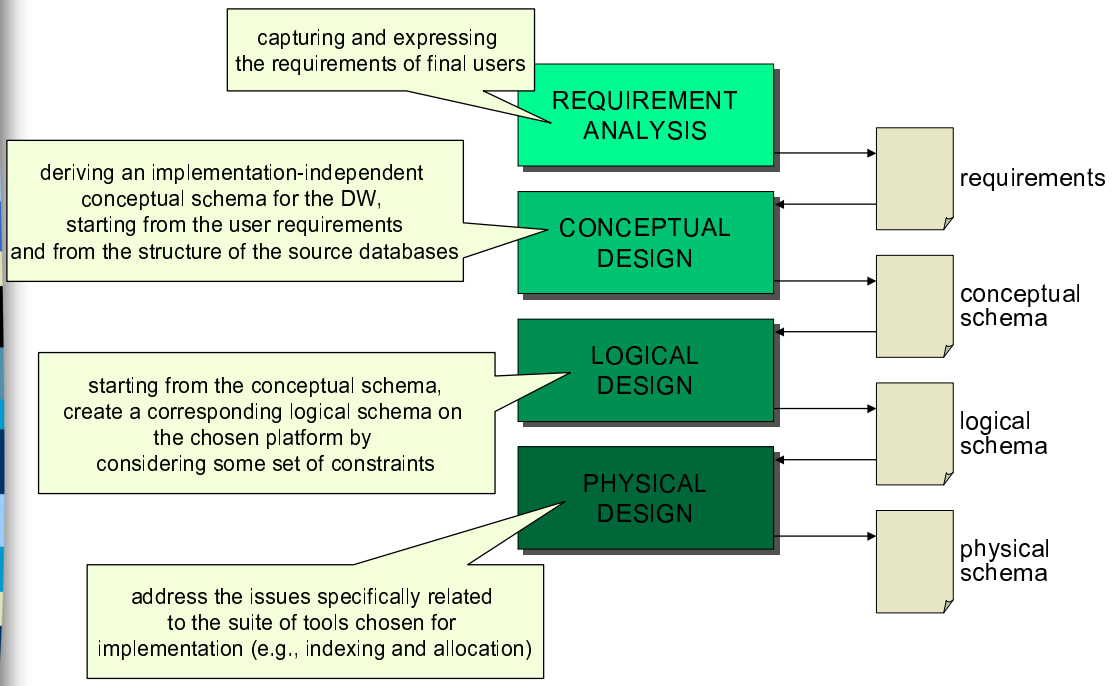
Research in Data Warehouse Modeling and Design: Dead or Alive?

Stefano Rizzi (Univ. of Bologna - Italy)
Alberto Abelló (Polytechnical Univ. of Catalunya - Spain)
Jens Lechtenbörger (Univ. of Münster - Germany)
Juan Trujillo (Univ. of Alicante - Spain)

Motivation

- This work follows a wide discussion that took place in Dagstuhl, during the Perspectives Workshop "Data Warehousing at the Crossroads" (August 2004)
- The aim of the seminar was to discuss the current trends in data warehousing
- Here the question is: "Has research on modeling and design come to an end? If not, what's left to do?"

Reference framework

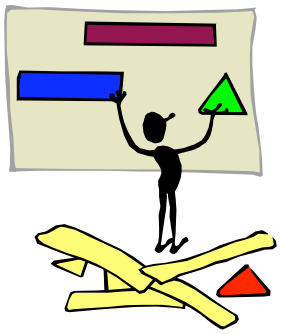


We focus on...

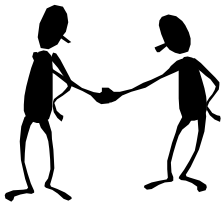
- Conceptual modeling
- Logical modeling
- Methods for design
- Interoperability and metadata
- Design for new architectures and applications

Conceptual modeling

- Multidimensional modeling
  - ✓ extensions to the Entity-Relationship model
  - ✓ extensions to UML
  - ✓ ad hoc models
- Modeling of ETL
  - ✓ from the functional point of view
  - ✓ from the dynamic point of view
  - ✓ from the static point of view

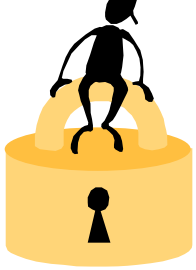


Issues: lack of a standard



- State of the art
  - ✓ still no agreement on the most relevant multidimensional properties to be modeled
  - ✓ some of the modeled properties cannot be expressed in the target logical models
  - ✓ no industrial push is given to any of the models
- Claim
  - ✓ A unified conceptual model for DWs, implemented within a CASE tools, would be a valuable support for both the research and industrial communities
- Challenges
  - ✓ it should be formally well-founded, easily usable and understandable by designers
  - ✓ it should support integrated modeling of the DW architecture, deployment, sources, mappings, ETL, facts, workloads, etc.
  - ✓ it should also support the peculiar issues and constraints arising in unusual and emerging domains and applications

Issues: modeling security



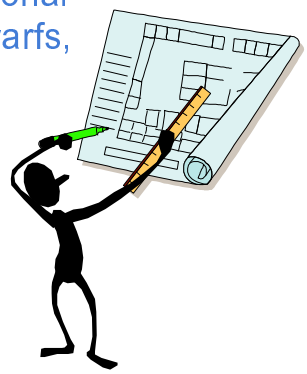
- State of the art
  - ✓ information security is a basic requirement for DWs
  - ✓ confidentiality is particularly relevant, since business information is sensitive
  - ✓ the security model used in transactional databases is unsuitable for DWs
- Claim
  - ✓ The classical security model should be replaced with a model centered on the concepts of multidimensional modeling and tightly integrated with the conceptual model
- Challenges
  - ✓ consider information security during all stages of the development life-cycle
  - ✓ consider all the components of the DW architecture, including ETL and sources
  - ✓ provide methods for transforming security models from the conceptual level into the logical level, and then into implementations
  - ✓ represent a hierarchy of roles for users, supported by a formal language to solve conflicts between different authorization rules

Issues: mining-aware design

- State of the art
  - ✓ some vendors already mix OLAP and data mining in tools
  - ✓ integrating OLAP and data mining as not been considered a hot research topic so far
  - ✓ DW design is mainly targeted at designing OLAP cubes, and no attention has been paid to consider mining requirements from the early stages of design
- Claim
  - ✓ Mining-aware design techniques and models should be devised
- Challenges
  - ✓ how could mining results be gracefully incorporated into DWs?
  - ✓ how could DW and OLAP storage techniques support data mining algorithms by facilitating them in accessing large volumes of cleansed and integrated data?
  - ✓ how would the two analysis techniques complement one another?

Logical modeling

- Considerable progress has been made in the area of multidimensional modeling, where target database systems are typically either relational or multidimensional
  - ✓ ROLAP: star, constellation, and snowflake schemata are widely accepted and supported by vendors
  - ✓ MOLAP: several efficient multidimensional data structures (condensed cubes, dwarfs, QC-Trees) have been proposed



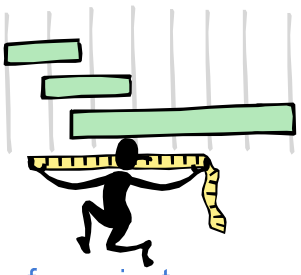
Issues: semantic gap

- State of the art
  - ✓ there is a semantic gap between advanced conceptual data models and (relational or multidimensional) implementations of data cubes
  - ✓ difficult to cope with generalization relationships in OLAP hierarchies
  - ✓ difficult to represent dimension constraints
  - ✓ difficult to treat summarizability using general aggregate functions
- Claim
  - ✓ This semantic gap should be bridged by preserving all information captured by conceptual multidimensional models in logical implementations
- Challenges
  - ✓ how to enrich meta-data for tool support in a systematic way?
  - ✓ look for more expressive logical models while preserving good query performance

Methods for design

- Several techniques for automating single phases of DW design have been proposed, but a very few comprehensive design methods have been devised so far
- In general, mechanisms should appear to coordinate all DW design phases allowing the analysis, control, and traceability of data and metadata along the project life-cycle (e.g., based on the Model Driven Architecture)

Issues: quality metrics

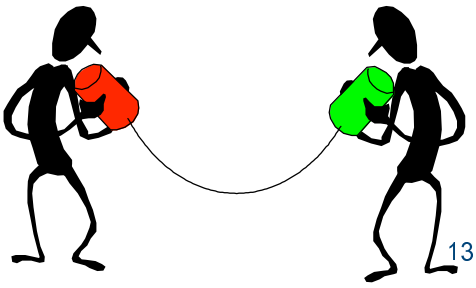


- State of the art
  - ✓ guarantee the DW quality from the early stages of a project
  - ✓ still no agreement on the quality of the design process and its impact on decision making
  - ✓ existing approaches at either the conceptual or the logical level
- Claim
  - ✓ More comprehensive metrics should be devised for measuring schema and data quality. They will support the designer in ranking different design alternatives, and will be useful to better plan the project and meet requirements
- Challenges
  - ✓ address the traceability of metrics and define thresholds to discriminate "good" schemata from "bad" ones
  - ✓ monitor the metrics and appropriately respond to their deviations during the DW lifetime
  - ✓ propagate data quality metrics to query results, and have data retrieval driven by the quality requirements expressed by users

Interoperability and metadata

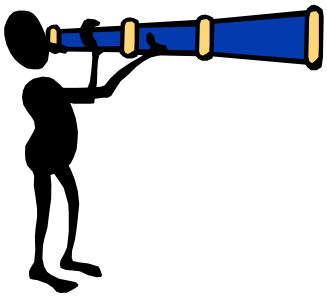
- State of the art
  - ✓ Broad diversity in metadata modeling in commercial tools
  - ✓ Tools are integrated by building complex metadata bridges, but some information is lost in translation
  - ✓ Two industry standards...
    - Open Information Model (OIM)
    - Common Warehouse Metamodel (CWM)

  ... but their expressivity is not sufficient to capture all the semantics represented by conceptual models
- Claim
  - ✓ There is a need for a standard definition of metadata in order to better support DW interoperability and integration



Design for new architectures and applications

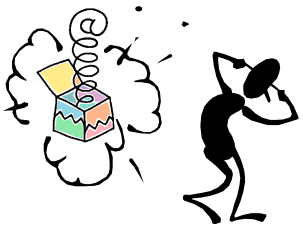
- The modeling and design techniques devised so far are mainly targeted towards traditional business applications, and aimed at managing simple alphanumerical data
- Advanced architectures for business intelligence are emerging to support new kinds of applications, possibly involving new and more complex data types
- Inevitably, more general techniques will have to be devised



Issues: spatial DWs

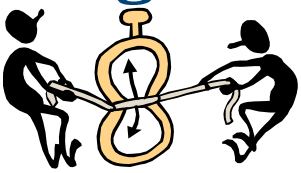
- State of the art
  - ✓ multidimensional models are extended with spatial dimensions, spatial hierarchies, spatial measures
  - ✓ topological relationships and operators such as intersect and inside as well as user-defined aggregate functions are included to augment the expressivity of these models
- Challenges
  - ✓ find more expressive conceptual representation for location data and devise adequate solutions for logical design in presence of spatial information
  - ✓ how to seamlessly integrate ROLAP and MOLAP solutions with the specialized data structures used in GISs while preserving high-level performance?

Issues: web warehousing



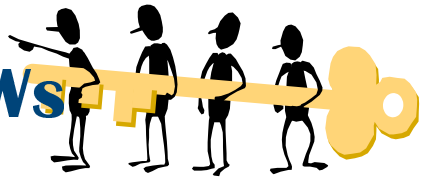
- State of the art
  - ✓ approaches for building a conceptual schema from XML data
  - ✓ approaches for driving design of the web warehouse by frequent user queries and data quality
- Claim
  - ✓ The development of the Semantic Web opens new exciting possibilities since knowledge is represented according to formal ontologies capable of expressing semantic relationships
- Challenges
  - ✓ how to integrate heterogeneous web sources?
  - ✓ how to automate conceptual design when some or most data sources reside on the Web?

Issues: real-time warehousing and BPM



- State of the art
  - ✓ some tools claim to provide BPM capabilities
  - ✓ some research on real-time ETL
- Challenges
  - ✓ achieving satisfactory performance for continuous monitoring queries requires more sophisticated logical models for data cubes
  - ✓ strict real-time will not be needed for most applications, but what is the *right-time* for the specific business domain?
  - ✓ devise suitable techniques for modeling and designing KPIs and business rules
  - ✓ understand business processes and their relationships in order to find out the relevant KPIs and rules, and to determine where the data to compute them can be found

Issues: distributed DWs



- State of the art
  - ✓ some research on adopting P2P infrastructures for warehousing XML resources
  - ✓ some research on deploying DWs on a grid
  - ✓ some approaches to fragmentation of DWs
- Challenges
  - ✓ how to design distribution?
  - ✓ how to deploy the DWs on the distributed architecture?
  - ✓ how to integrate heterogeneous data marts?

Conclusion

- Though DW modeling and design have been investigated for about a decade, several important challenges still arise
  - ✓ Ad hoc techniques are required for dealing with the emerging applications of data warehousing and with advanced architectures for business intelligence
  - ✓ The need for real-time data processing raises original issues that were not addressed within traditional periodically-refreshed DWs