Data warehouse Introduction

Elena Baralis
Politecnico di Torino

Decision support systems

- Huge operational databases are available in most companies
  - ⇨these databases may provide a large wealth of useful information
- Decision support systems provide means for
  - ⇨in depth analysis of a company's business
  - ⇨*faster* and *better* decisions

Strategic decision support

- Demand evolution analysis and forecast
- Critical business areas identification
- Budgeting and management transparency
  - reporting, practices against frauds and money laundering
- Identification and implementation of winning strategies
  - ⇨cost reduction and profit increase

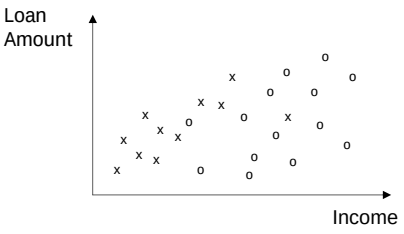
Business Intelligence

- BI provides support to strategic decision support in companies
- Objective: transforming company data into actionable information
  - at different detail levels
  - for analysis applications
- Users may have heterogeneous needs
- BI requires an appropriate hardware and software infrastructure

Applications

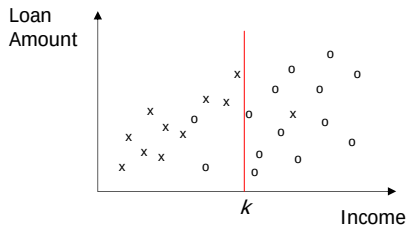
- Manifacturing companies: order management, client support
- Distribution: user profile, stock management
- Financial services: buyer behavior (credit cards)
- Insurance: claim analysis, fraud detection
- Telecommunication: call analysis, churning, fraud detection
- Public service: usage analysis
- Health: service analysis and evaluation
- ... and many more...

Example



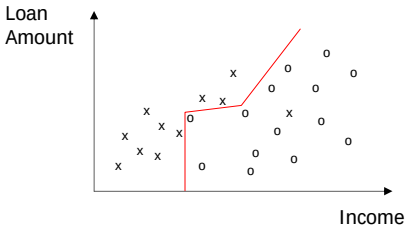
Bank clients with a loan
x: "bad" clients owing periodic payments to the bank after due date
o: "good" clients respecting periodic payment due date

Example



If Income < *k€* **then** "bad" client

Example



Data management

- Traditional DBMS usage, characterized by
  - detailed data, relational representation
  - snapshot of current data state
  - well-known, structured and repetitive operations
  - read/write access to few records
  - short transactions
  - isolation, reliability and integrity (ACID) are critical
  - database size ≈ 100MB-GB

Data analysis

- Data processing for decision support, characterized by
  - "historical" data
  - consolidated and integrated data
  - ad hoc applications
  - read access to millions of record
  - complex queries
  - data consistency before and after periodical loads
  - database size ≈ 100GB-TB

Data warehouse

- Database devoted to decision support, which is kept *separate* from company operational databases
- Data which is
  - integrated
  - time dependent, non volatile
  - devoted to a specific subject

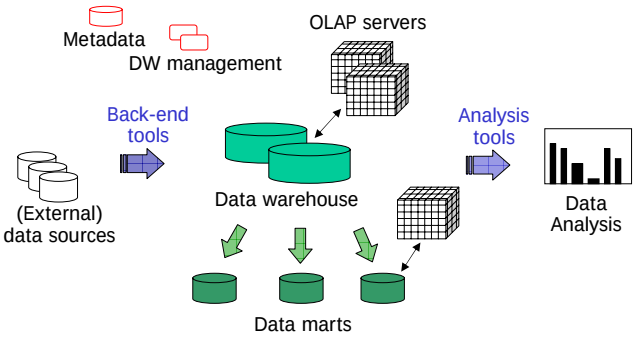
  used for decision support in a company

W. H. Inmon, Building the data warehouse, 1992

Why separate data?

- Performance
  - complex queries reduce performance of operational transaction management
  - different access methods at the physical level
- Data management
  - missing information (e.g., history)
  - data consolidation
  - data quality (inconsistency problems)

Data warehouse: architecture



Data warehouse and data mart

Company data warehouse: it contains *all* the information on the company business
- extensive functional modelling process
- design and implementation require a long time

Data mart: departimental information subset focused on a given subject
- two architectures
  - dependent, fed by the company data warehouse
  - independent, fed directly by the sources
- faster implementation
- requires careful design, to avoid subsequent data mart integration problems

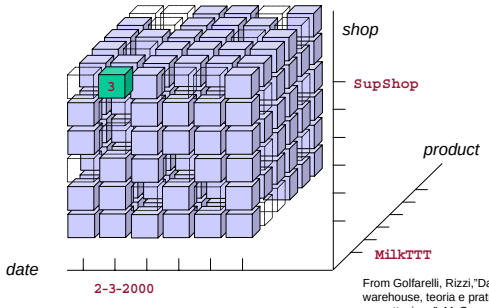
Back-end tools

- Feed the data warehouse
  (ETL = Extraction Transformation Loading)
  - data extraction from data sources
  - data cleaning (errors, missing or duplicated data)
  - format trasformations and conversions
  - data loading and periodical refresh

Multidimensional representation

- Data are represented as an (hyper)cube with three or more dimensions
- Measures on which analysis is performed: cells at dimension intersection
- Data warehouse for tracking sales in a supermarket chain:
  - dimensions: product, shop, time
  - measures: sold quantity, sold amount, ...

Multidimensional representation



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

Data analysis tools

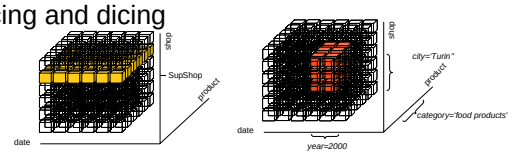
- OLAP analysis: complex aggregate function computation
  - support to different types of aggregate functions (e.g., moving average, top ten)
- Data analysis by means of data mining techniques
  - various analysis types
  - significant algorithmic contribution

Data analysis tools

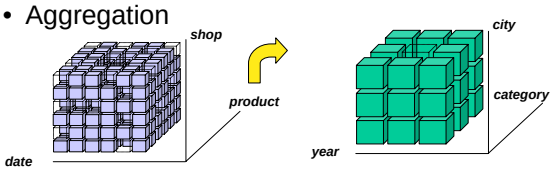
- Presentation
  - separate activity: data returned by a query may be rendered by means of different presentation tools
- Motivation search
  - Data exploration by means of progressive, "incremental" refinements (e.g., drill down)

OLAP analysis

- Slicing and dicing



- Aggregation



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

Types of data mining activities

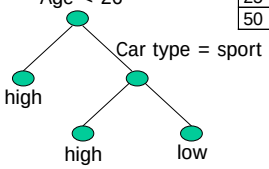
Classification and regression: predictive model generation
- requires a previously labeled data set

Association rules: extraction of data correlations

Clustering: data partioned in "homogeneous" groups
- requires the notion of distance between two elements

Example: classification

Age	Car type	Risk category
40	SW	low
65	sport	high
20	utility	high
25	sport	high
50	utility	low



Decision tree

Example: association rules

- Given a collection of counter transactions in a supermarket (receipts)
- Association rules

$$\text{diapers} \Rightarrow \text{beer}$$

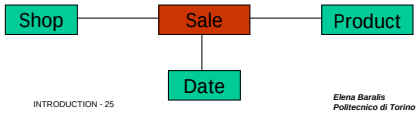
  - 2% of transactions contains both elements
  - 30% of transactions containing diapers also contains beer

Servers for Data Warehouses

- ROLAP (Relational OLAP) server
  - extended relational DBMS
    - compact representation for sparse data
  - SQL extensions for aggregate computation
  - specialized access methods which implement efficient OLAP data access
- MOLAP (Multidimensional OLAP) server
  - data represented in proprietary (multidimensional) matrix format
    - sparse data require compression
  - special OLAP primitives
- HOLAP (Hybrid OLAP) server

Relational representation: star model

- (Numerical) measures stored in the *fact table*
  - attribute domain is numeric
- *Dimensions* describe the context of each measure in the fact table
  - characterized by many descriptive attributes
- Example: Data warehouse for tracking sales in a supermarket chain



Data warehouse size

- Time dimension: 2 years x 365 days
- Shop dimension: 300 shops
- Product dimension: 30.000 products, of which 3.000 sold every day in every shop
- Number of rows in the fact table:

  730 x 300 x 3000 = 657 millions

⇨ Size of the fact table ≈ 21GB

Meta data

- Different types of meta data:
  - for data transformation and loading:
    - describe data sources and needed transformation operations
  - for data management:
    - describe the structure of the data in the data warehouse (also for materialized view)
  - for query management:
    - data on query structure and execution
      - SQL code for the query
      - execution plan
      - memory and CPU usage

Textbooks

- Data warehousing

  Golfarelli, Rizzi, *Data warehouse: teoria e pratica della progettazione*, McGraw-Hill 2006

  Kimbal et al., textbooks on metodology and case studies, Wiley
- Data mining

  Han, Kamber, *Data mining: concepts and techniques*, Morgan Kaufmann 2006

  Tan, Steinbach, Kumar, *Introduction to data mining*, Pearson 2006

Useful links

- Data warehouse

  http://www.dwinfocenter.org

  http://www.dwreview.com

  http://kimballuniversity.com
- Data mining

  http://www.kdnuggets.com/