Data Warehousing

Parte III

Data mining e Progetto

Argomenti della lezione

- Data mining
- Progetto di un data warehouse
- Integrazione di sorgenti informative
- Reverse Engineering
- Esempio di progetto

Data mining

- Approccio alternativo all'analisi multidimensionale per estrarre informazioni di supporto alle decisioni da un data warehouse
- Insieme di tecniche di ricerca di "informazione nascosta" in una collezione di dati

Problemi di data mining

- associazioni: individuare regolarità in un insieme di transazioni anonime
- pattern sequenziali: individuare regolarità in un insieme di transazioni non anonime, nel corso di un periodo temporale
- classificazione: catalogare un fenomeno in una classe predefinita sulla base di fenomeni già catalogati

Associazioni

Dati di ingresso:

- sequenze di oggetti (transazioni)

Obiettivo:

- trovare delle "regole" che correlano la presenza di un insieme di oggetti con un altro insieme di oggetti

Esempio di regola

Pannolini ⇒ Birra

- il 30% delle transazioni che contiene Pannolini contiene anche Birra
- il 2% tra tutte le transazioni contiene entrambi gli oggetti

Rilevanza delle regole

$$X, Y \Rightarrow Z$$

- **Supporto S: la regola è verificata in S% delle transazioni rispetto a tutte le transazioni**
  - rilevanza statistica
- **Confidenza C: C% di tutte le transazioni che contengono X e Y contengono anche Z**
  - "forza" della regola

Pattern sequenziali

Dati di ingresso:

- insieme di transazioni eseguita da un certo cliente

Obiettivo:

- trovare le sequenze di oggetti che compaiono in almeno una data percentuale data di insiemi di transazioni

Esempi

- il 5% dei clienti ha comprato un lettore di CD in una transazione e CD in un'altra
  - il 5% è il supporto del pattern
- Applicazioni
  - misura della soddisfazione del cliente
  - promozioni mirate
  - medicina (sintomi - malattia)

Progettazione del warehouse

La progettazione di un data warehouse è diversa dalla progettazione di una base di dati operazionale

- i dati da memorizzare hanno caratteristiche diverse
- vincolata dalle basi di dati esistenti
- guidata da criteri progettuali diversi

Attività principali

- analisi
  - requisiti
  - sorgenti informative esistenti
- integrazione
- progettazione
  - concettuale
  - logica
  - fisica


Requisiti di analisi
Sorgenti informative
Analisi
Schemi sorgenti informative
Integrazione
Schema integrato
Progettazione
Schema del DW

Dati in ingresso

Le informazioni in ingresso necessarie alla progettazione di un data warehouse:

- requisiti di analisi
- descrizione delle basi di dati operazionali disponibili
- descrizione di altre sorgenti informative esteme

Fase di analisi

- Selezione delle sorgenti informative
  - analisi preliminare del patrimonio informativo aziendale
  - correlazione con i requisiti
  - identificazione di priorità tra schemi
- Traduzione in un modello concettuale di riferimento
- Analisi delle sorgenti informative
  - identificazione di fatti, misure e dimensioni

Fase di integrazione

fusione dei dati rappresentati in più sorgenti in un'unica base di dati globale

- identificazione di concetti comuni
- unificazione mediante risoluzione di conflitti
  - terminologici
  - strutturali
  - di codifica

Fase di progettazione

- concettuale: completare la rappresentazione dei concetti dimensionali necessari per l'analisi
  - ad esempio, dati storici e geografici
- logica: identificare il miglior compromesso tra la necessità di aggregare i dati e quella di normalizzarli
- fisica: individuare la distribuzione dei dati e le relative strutture di accesso

Reverse engineering

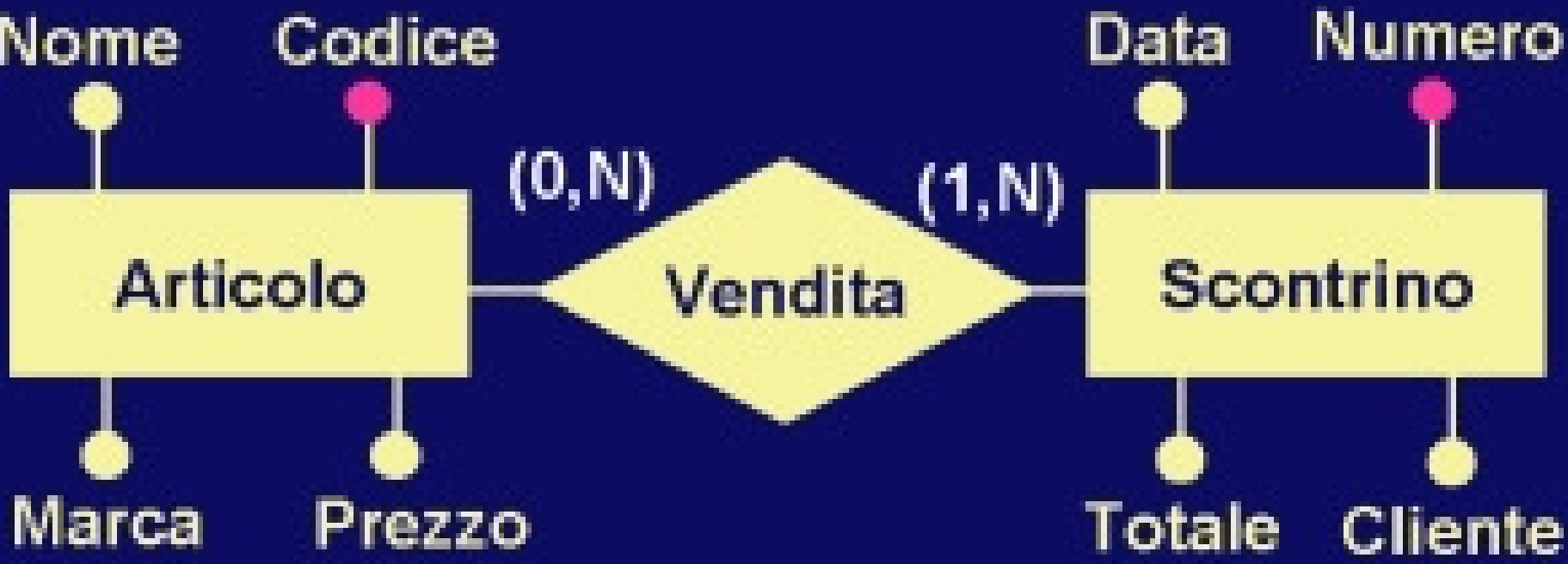
- Attività della fase di analisi
- Rappresentazione concettuale di uno schema di base di dati
- Il reverse engineering di schemi relazionali è svolto in modo semiautomatico dagli strumenti di progettazione CASE

Esempio

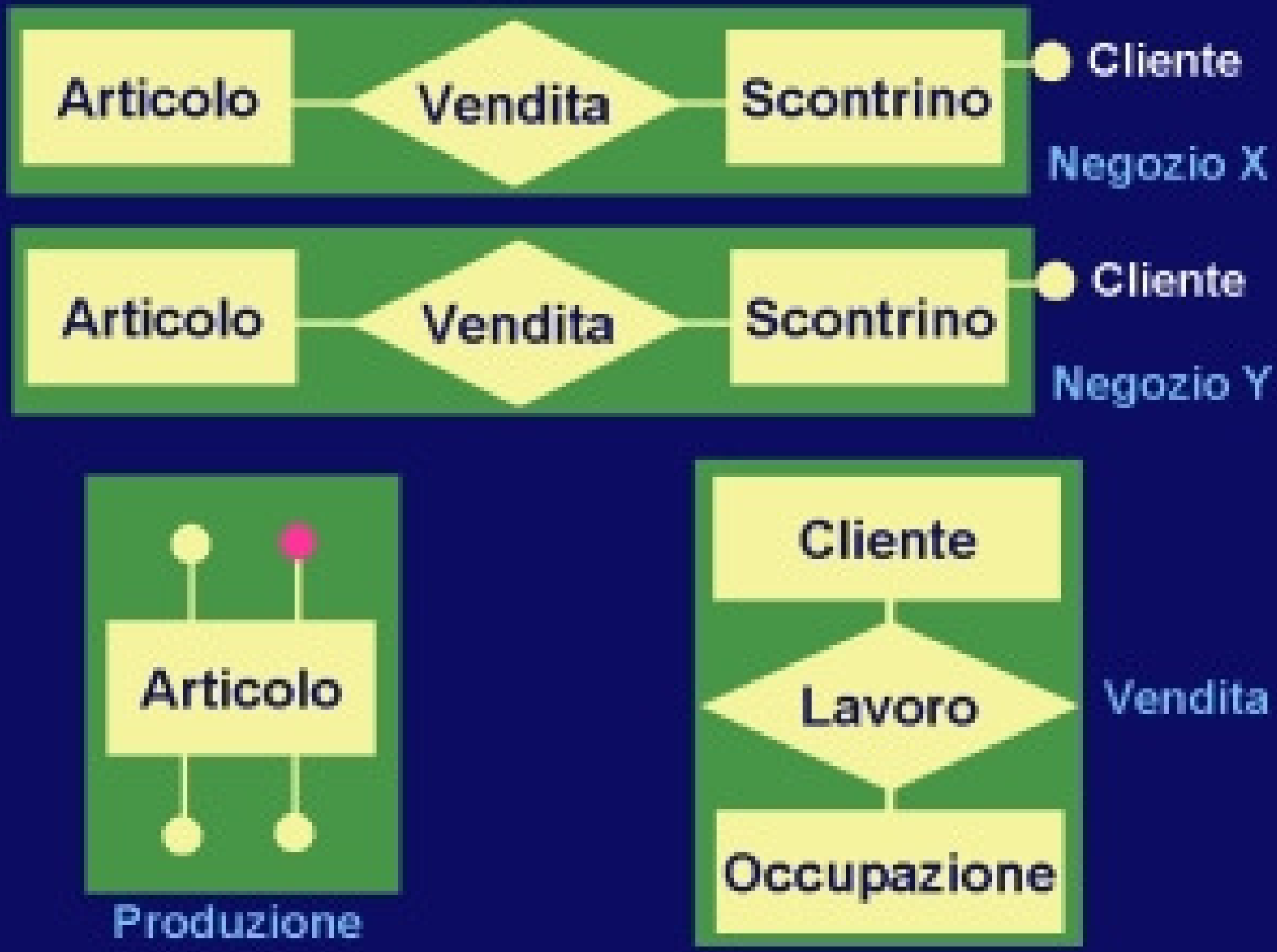
Articolo (Codice, Nome, Prezzo, Marca)

Scontrino (Numero, Data, Totale, Cliente)

Vendita (Articolo, Scontrino)



Esempio di integrazione di schemi


Articolo
Vendita
Scontrino
Cliente
Negozio X
Articolo
Vendita
Scontrino
Cliente
Negozio Y
Articolo
Produzione
Cliente
Lavoro
Vendita
Occupazione


Articolo
Vendita
Scontrino
Negozio
Spesa
Dati
Cliente
Dettaglio
Lavoro
Occupazione

Integrazione di istanze

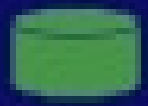
È guidata da quella dei loro schemi ma è necessario risolvere i conflitti

- ad esempio, un attributo “sesso” può essere rappresentato
  - con un singolo carattere — M / F
  - con una singola cifra — 0 / 1
  - con un valore logico — vero / falso
  - implicitamente nel codice fiscale
  - non essere rappresentato

Problema legato alla qualità dei dati disponibili

Mario Rossi è nato il 1 marzo 1942

Mario Rossi è nato il 10 marzo 1942



Mairo Rossi è nato il 10 marzo 1942

Progettazione del DW

Introduzione di elementi dimensionali nella base di dati integrata

- identificazione di fatti, misure, dimensioni
- ristrutturazione
  - rappresentazione di fatti mediante entità
  - individuazione di nuove dimensioni
  - raffinamento dei livelli di ogni dimensione
- progettazione logica e fisica

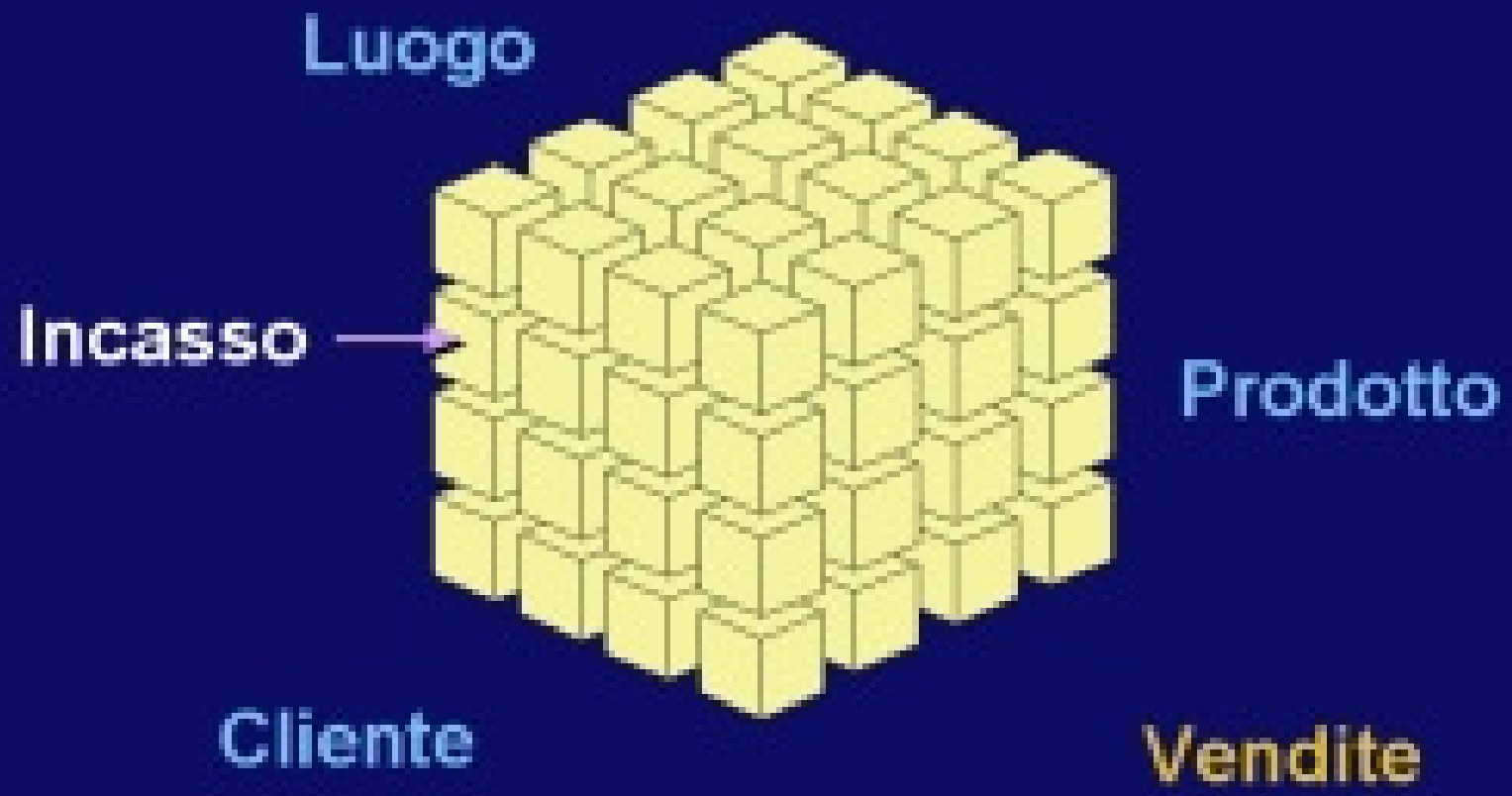

Articolo
Prodotto
Vendita
Incasso
Dati
Spesa
M
Cliente
Negozio
Dettaglio
Lavoro
A
Occupazione
Regione

Progettazione logica ROALP

Progettazione logica MOLAP

Si costruiscono delle matrici a n dimensioni le cui celle contengono i dati

Le gerarchie sui livelli sono codificate in indici di accesso alle matrici


Luogo
Incasso
Prodotto
Cliente
Vendite

Progettazione fisica

- Definizione di indici di accesso
- Definizione di viste materializzate
- Definizione di pre-aggregazioni

Estensioni della metodologia

La metodologia può essere estesa per includere sorgenti informative che risiedono su internet

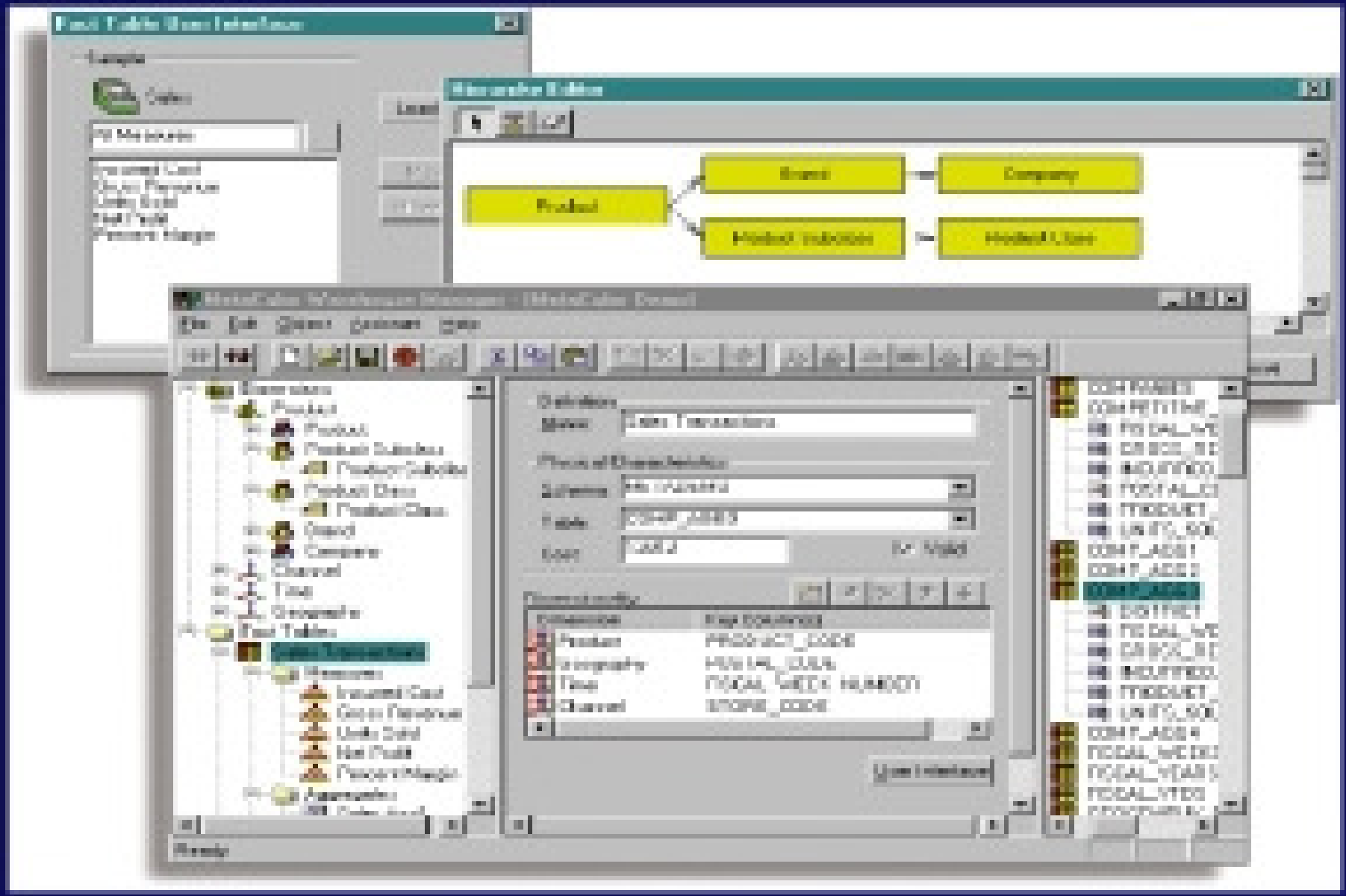
Attività ulteriori (nell'analisi):

- discovering
- costruzione di wrapper
- strutturazione
- caricamento

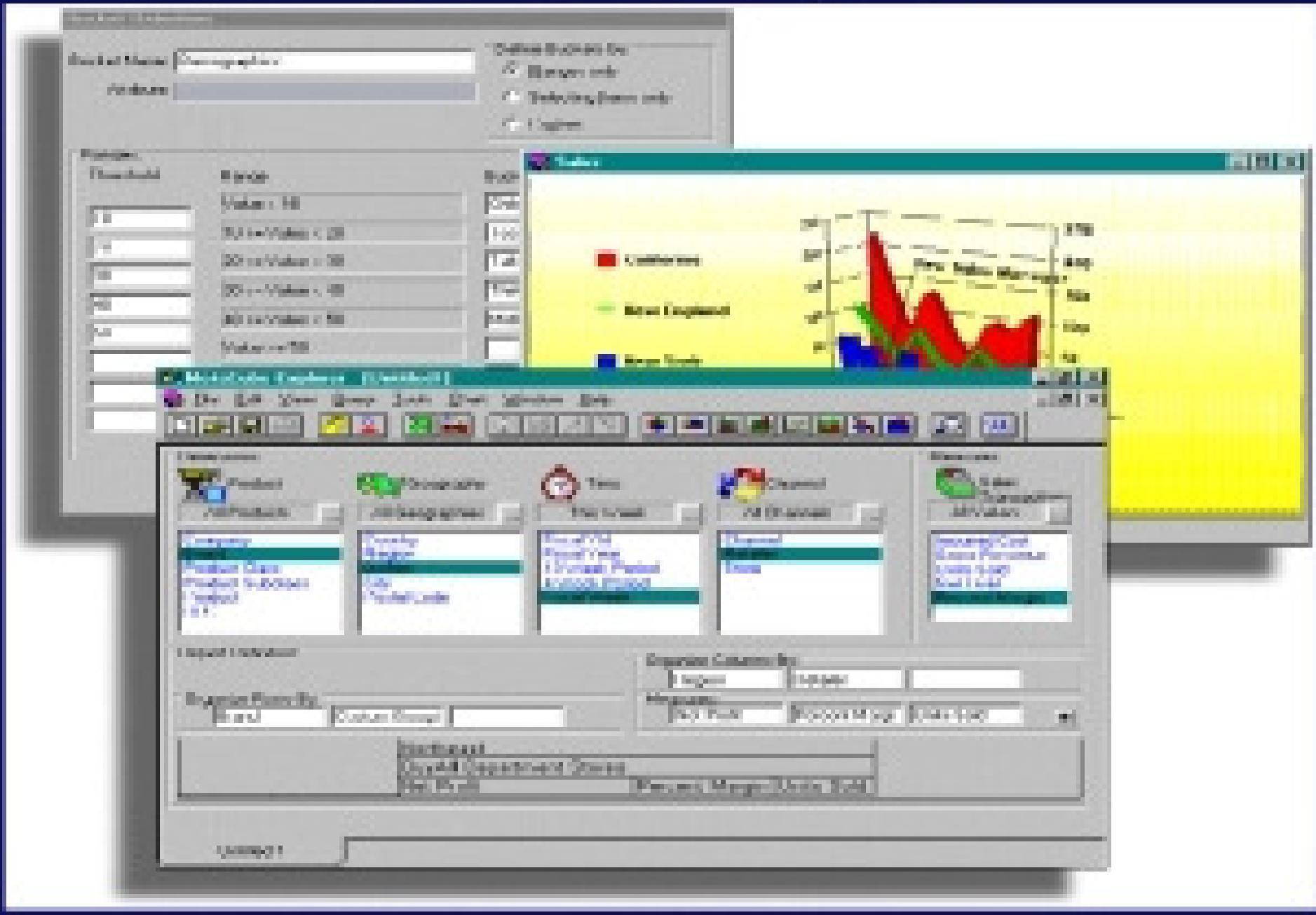
Considerazioni finali

- Molti passi della metodologia possono essere automatizzati
- Alcuni passi si basano su metodi e tecniche note (quadro metodologico)
- Alcune fasi sono valide a prescindere dalla realizzazione di un data warehouse
- Esistono strumenti CASE di supporto

MetaCube Warehouse Manager



MetaCube Explorer



Sommario

- Data mining
- Progetto di un data warehouse
- Integrazione di sorgenti informative
- Reverse Engineering
- Esempio di progetto