Data warehouse

Data warehouse

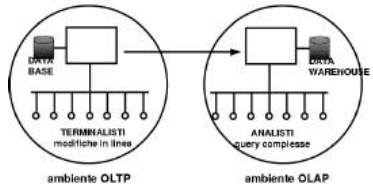
- La crescita dell'importanza dell'analisi dei dati ha portato ad una separazione architetturale dell'ambiente transazionale (OLTP *on-line transaction processing*) da quello dedicato all'analisi (OPLAP *on-line analytical processing*).
- Conseguenza è nata una classe di sistemi detta data warehouse (magazzini di dati), dedicati all'elaborazione e analisi dei dati.

Data warehouse

Definizione:

L'insieme delle strutture dati e dei tool necessari per ottenere, a partire dai dati operazionali utilizzati e creati dal sistema informativo aziendale, informazioni che aiutino i manager nella valutazione tecnico-economica dell'andamento aziendale

Architettura complessiva con OLTP e OLAP



OLTP

• Tradizionale elaborazione di transazioni, che realizzano i processi operativi dell'azienda-ente

- Operazioni predefinite e relativamente semplici
- Ogni operazione coinvolge "pochi" dati
- Dati di dettaglio, aggiornati
- Le proprietà "acide" (atomicità, correttezza, isolamento, durabilità) delle transazioni sono essenziali

Sistemi di supporto alle decisioni

• Richiedono operazioni non previste a priori

• Coinvolgono spesso grandi quantità di dati, anche storici e aggregati

• Coinvolgono dati provenienti da varie fonti operative, anche esterne

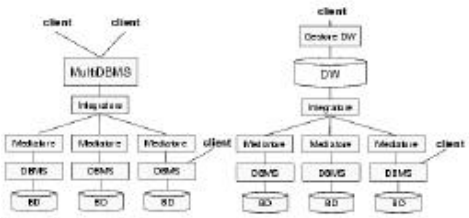
OLAP

- **Elaborazione di operazioni per il supporto alle decisioni:**
  - Operazioni complesse e casuali
  - Ogni operazione può coinvolgere molti dati
  - Dati aggregati, storici, anche non attualissimi
  - Le proprietà “acide” non sono rilevanti, perché le operazioni sono di sola lettura

OLAP e OLTP

- I requisiti sono quindi contrastanti
- Le applicazioni dei due tipi possono danneggiarsi a vicenda

Multi-database e data warehouse



Data warehouse

Una base di dati:

- utilizzata principalmente per il supporto alle decisioni direzionali
- integrata verso la realtà aziendale e non dipartimentale
- orientata ai dati e non alle applicazioni
- dati sono storici con un ampio orizzonte temporale, e indicazione (di solito) di elementi di tempo
- Dati non volatile i dati sono caricati e acceduti fuori linea
- Base di dati mantenuta separatamente dalle basi di dati operazionali

DW: integrata

- I dati di interesse provengono da tutte le sorgenti informative: ciascun dato proviene da una o più di esse

- Il data warehouse rappresenta i dati in modo univoco : riconciliando le eterogeneità dalle diverse rappresentazioni
  - nomi
  - codifica
  - rappresentazione multipla

DW: orientata ai dati

- Le basi di dati operazionali sono costruite a supporto dei singoli processi operativi o applicazioni
  - produzione
  - vendita

- Il data warehouse è costruito attorno alle principali entità del patrimonio informativo aziendale
  - prodotto
  - cliente

DW: dati storici

- Le basi di dati operazionali mantengono il valore corrente delle informazioni
- L'orizzonte temporale di interesse è dell'ordine dei pochi mesi
- Nel data warehouse è di interesse l'evoluzione storica delle informazioni
- L'orizzonte temporale di interesse è dell'ordine degli anni

DW: non volatile

- In una base di dati operazionale, i dati vengono
  - acceduti, inseriti, modificati, cancellati
- pochi record alla volta
- Nel data warehouse, abbiamo
  - operazioni di accesso e interrogazione - "diurne"
  - operazioni di caricamento e aggiornamento dei dati - "notturne"
- che riguardano milioni di record

DW: una base di dati separata

- Per tanti motivi
  - non esiste un'unica base di dati operazionale contenente tutti i dati di interesse
  - la base di dati deve essere integrata
  - non è tecnicamente possibile fare l'integrazione in linea
  - i dati di interesse sarebbero comunque diversi
- devono essere mantenuti dati storici
- devono essere mantenuti dati aggregati
  - l'analisi dei dati richiede per i dati organizzazioni speciali e metodi di accesso specifici
  - degrado generale delle prestazioni senza la separazione

Popolazione del data warehouse

- Attività neccessarie:
  - estrazione
  - trasformazione
  - caricamento
  - refresh
- I ***metadati*** sono informazioni mantenute a supporto di queste attività

Livelli di rappresentazione dei dati

- Nelle **sorgenti informative**
  - dipartimentali: orientate alle applicazioni; es: vendita, produzione, marketing, ...
- Nel **data warehouse**
  - aziendale: soggetti comuni es: prodotti, clienti, fornitori, ...

Livelli di rappresentazione dei dati

- Nei **data mart**
  - dipartimentali o settoriali: selezionati per un particolare problema
    es: dati relativi al marketing
- Negli **strumenti di analisi**
  - individuali: focalizzata su un problema in esame
    es: vendite negli ultimi cinque anni

Differenza fra data warehouse e DBMS tradizionale

- Modalità d'uso
  - funzionamento "normale": query di sola lettura
  - aggiornamento: lunghi programmi batch
- Progetto fisico dei dati
  - supporto accesso sequenziale
  - indici sparsi, strutture dati "invertite"
  - clusterizzazioni e raggruppamenti predefiniti
  - parallelismo intra-query
- Ambiente di sviluppo
  - orientato ad utenti non informatici

Problemi di progetto

- Definizione dei requisiti
  - individuazione dei principali problemi decisionali per l'impresa
- Scelta dei dati
  - individuazione del "data source" coinvolte
- Definizione delle modalità di acquisizione
  - uso del "replication manager"
- Miglioramento della qualità dei dati
  - filtro di dati scorretti
  - integrazione di dati da plurime fonti

Modelli e linguaggi per l'analisi dei dati

Il problema

- Limitazioni della tecnologia relazionale
  - difficoltà d'uso
  - rigidità
- Conseguenze
  - uso operativo: buono
  - uso strategico: scarso
- Reazione:
  - modelli, linguaggi tecniche per on-line analytical processing (OLAP)

Obiettivi di OLAP

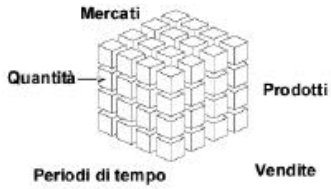
- Definire una versione consistente, pubblica, di qualità dei dati aziendali
- Facilitare l'accesso ai dati per uso strategico
- Applicazioni di OLAP:
  - supporto alle decisioni e business planning (finanze, marketing, vendite)

Modello multidimensionale

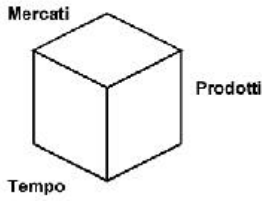
- E' una struttura entità-relazioni semplificata su cui fare interrogazioni standard
- Database strutturato in :
  - fatti
  - dimensioni di analisi

Data cube: aggregazione in SQL che permette di esprimere tute le aggregazioni possibili delle tuple (righe) di una tabella

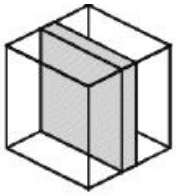
Rappresentazione multidimensionale dei dati



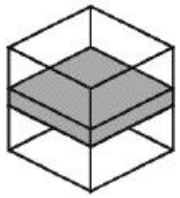
Viste sui dati multidimensionali



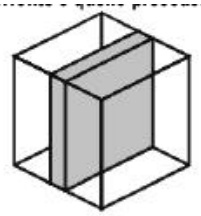
Il manager regionale esamina la vendita dei prodotti in tutti i periodi relativamente ai propri mercati



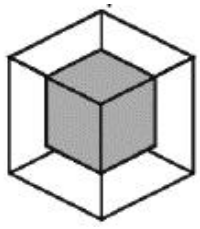
Il manager di prodotto esamina la vendita di un prodotto in tutti i periodi e in tutti i mercati



Il manager finanziario esamina la vendita dei prodotti in tutti i mercati relativamente al periodo corrente e quello precedente

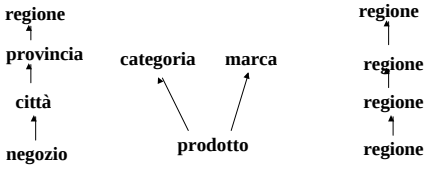


Il manager strategico si concentra su una categoria di prodotti, una area regionale e un orizzonte temporale medio



Dimensioni e gerarchie di livelli

- Ciascuna dimensione è organizzata in una gerarchia che rappresenta i possibili livelli di aggregazione per i dati



Operazioni su dati multidimensionali

- **Roll up** : aggrega i dati
  - volume di vendita totale dello scorso anno per categoria di prodotto e regione
- **Drill down** : disaggrega i dati
  - per una particolare categoria di prodotto e regione, mostra le vendite giornaliere dettagliate per ciascun negozio
- **Slice & dice** : seleziona e proietta
- **Pivot** : re-orienta il cubo

Visualizzazione dei dati

- I dati vengono visualizzati in veste grafica, in maniera da essere facilmente comprensibili
- Si fa uso di:
  - tabelle, istogrammi, grafici, torte, superfici 3D, bolle, area in pila, etc...

Progettazione di data 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 delle sorgenti informative esistenti
  - integrazione
  - progettazione concettuale, logica e fisica

Progettazione di data warehouse

- **Input**
  - requisiti dell'analisi
  - basi di dati aziendali
  - Altre sorgenti informative
- **Analisi**
  - selezione delle sorgenti informative
  - traduzione in un modello concettuale comune
  - Analisi delle sorgenti informative
- **Integrazione**
  - integrazione di schemi concettuali
- **Progettazione**
  - progettazione concettuale
  - progettazione logica
  - progettazione fisica

Data mining

- ***Obiettivo:***

  estrarre informazione nascosta nei dati in modo da consentire decisioni strategiche
- Una materia interdisciplinare

  statistica, algoritmica, reti neurali, geometria frattale

Applicazioni del data mining

- Analisi di mercato
  - prodotti acquisiti insieme o in sequenza
- Analisi di comportamento
  - individuare usi illeciti di creditcard
- Previsione
  - prevedere il costo delle cure mediche
- Controllo
  - errori di produzione

Regole di associazione

- Ricercano regolarità nei dati:
  - quando si acquistano scarponi, si acquistano sci
- Strutturate come:
  - corpo: premessa della regola
  - testa: conseguenza della regola

Caratteristiche delle regole di associazione

- Supporto
  - probabilità che siano presenti in una transazione entrambi gli elementi di una regola
- Confidenza
  - probabilità che sia presente in una transazione la testa di una regola, essendo presente il corpo
- Formulazione del problema
  - estrarre tutte le regole con supporto e confidenza superiori ai valori prefissati

Esempi di regole di associazione

Corpo	Testa	Supporto	Confidenza
pantaloni-sci	scarponi	0.25	1
scarponi	pantaloni-sci	0.25	1
magliette	stivali	0.25	0.5
magliette	giacche	0.25	1
stivali	magliette	0.25	0.5
stivali	giacche	0.25	1
giacche	magliette	0.5	0.66
giacche	stivali	0.25	0.33
{magliette,stivali}	giacche	0.25	1
{magliette,giacche}	stivali	0.25	0.5
{stivali, giacche}	magliette	0.25	1

Altri esempi

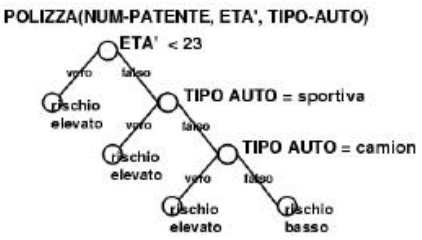
- Oggetti venduti nella stessa produzione
- Oggetti venduti assieme d'estate ma non d'inverno
- Oggetti venduti assieme in quanto disposti in modo particolare
- Oggetti acquisiti in sequenza dallo stesso cliente

Classificazione

- Catalogazione di un fenomeno particolare in una classe predefinita
  - fenomeno presentato sotto forma di fatti elementari (tupla)
  - Costruzione del classificatore a partire da un set di dati di prova (training set)
  - Classificatori rappresentati come alberi di decisione

Esempio di classificatore

individuazione di polizze a rischio



Sintesi dei vari aspetti presenti nella analisi dei dati

- Ambiente
  - data warehouse
- Modello
  - multidimensionali
- Estensioni di SQL
  - data cube
- Tecnologie di base
  - distribuzione
  - parallelismo
  - replicazione
- Tecnologie specifiche
  - browser (e visualizzatori)
  - data mining
    - associazione
    - classificazione