Data warehouse Introduzione

Elena Baralis
Politecnico di Torino

Copyright – Tutti i diritti riservati

Supporto alle decisioni aziendali

- La maggior parte delle aziende dispone di enormi basi di dati contenenti dati di tipo operativo
  - ⇨queste basi di dati costituiscono una potenziale miniera di informazioni utili
- Sistemi per il supporto alle decisioni permettono di
  - ⇨analizzare lo stato dell'azienda
  - ⇨prendere decisioni *rapide* e *migliori*

Copyright – Tutti i diritti riservati

Supporto alle decisioni aziendali

- Analisi e previsione dell'evoluzione della domanda
- Individuazione di aree critiche
- Chiarezza dei conti e trasparenza finanziaria
  - – reporting, pratiche antifrode e antiriciclaggio
- Definizione e realizzazione di strategie vincenti
  - ⇨contenimento di costi e aumento di profitti

Copyright – Tutti i diritti riservati

Business Intelligence

- Disciplina di supporto alla decisione strategica aziendale
- Obiettivo: trasformazione dei dati aziendali in informazioni fruibili
  - – a livelli diversi di dettaglio
  - – per applicazioni di analisi
- Tipologia di utenza eterogenea
- Necessaria un'adeguata infrastruttura hardware e software di supporto

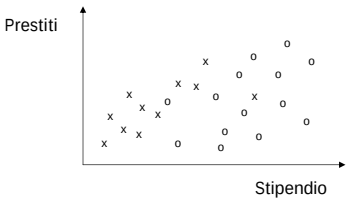
Copyright – Tutti i diritti riservati

Ambiti applicativi

- Industrie manifatturiere: gestione ordini e spedizioni, supporto clienti
- Distribuzione: profilo utenti, gestione magazzino
- Servizi finanziari: analisi acquisti (carta di credito)
- Assicurazioni: analisi richieste indennizzo, riconoscimento frodi
- Telecomunicazioni: analisi delle chiamate, riconoscimento frodi
- Servizi pubblici: analisi dell'utilizzo
- Sanità: analisi dei risultati

Copyright – Tutti i diritti riservati

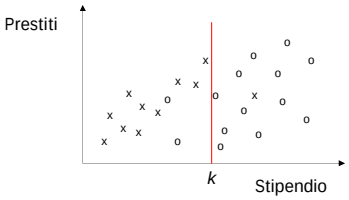
Esempio



Clienti di una banca che hanno contratto un prestito
x: clienti che hanno mancato la restituzione di rate
o: clienti che hanno rispettato le scadenze

Copyright – Tutti i diritti riservati

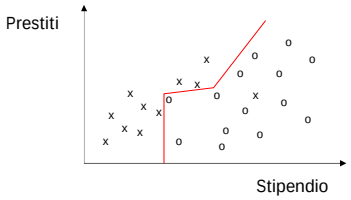
Esempio



If stipendio < *k*€ **then** mancati pagamenti

Copyright – Tutti i diritti riservati

Esempio



Copyright – Tutti i diritti riservati

Elaborazione dei dati

- Modalità tradizionale di uso dei DBMS, caratterizzata da
  - istantanea del valore corrente dei dati
  - dati dettagliati, rappresentazione relazionale
  - operazioni strutturate e ripetitive
  - accesso in lettura o aggiornamento di pochi record
  - transazioni brevi
  - isolamento, affidabilità e integrità sono critici
  - dimensione della base di dati ≈ 100MB-GB

Copyright – Tutti i diritti riservati

Analisi dei dati

- Elaborazione dei dati per il supporto alle decisioni, caratterizzata da
  - dati di tipo “storico”
  - dati consolidati e integrati
  - applicazioni ad hoc
  - accesso in lettura a milioni di record
  - interrogazioni di tipo complesso
  - consistenza dei dati prima e dopo le operazioni di caricamento periodico
  - dimensione della base di dati ≈ 100GB-TB

Copyright – Tutti i diritti riservati

Data warehouse

- Base di dati per il supporto alle decisioni, che è mantenuta *separatamente* dalle basi di dati operative dell’azienda
- Dati
  - orientati ai soggetti di interesse
  - integrati e consistenti
  - dipendenti dal tempo, non volatili

  utilizzati per il supporto alle decisioni aziendali

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

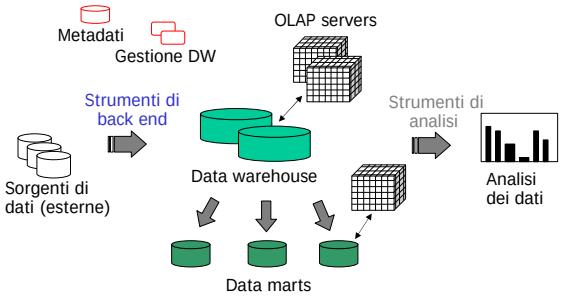
Copyright – Tutti i diritti riservati

Perché dati separati?

- Prestazioni
  - ricerche complesse riducono le prestazioni delle transazioni operative
  - metodi di accesso diversi a livello fisico
- Gestione dei dati
  - informazioni mancanti (storico)
  - consolidamento dei dati
  - qualità dei dati (problema di inconsistenze)

Copyright – Tutti i diritti riservati

Elementi costitutivi di un data warehouse



Data warehouse e data mart

Warehouse aziendale: contiene informazioni sul funzionamento di "tutta" l'azienda
- processo di modellazione funzionale esteso
- progettazione e realizzazione richiedono molto tempo

Data mart: sottoinsieme dipartimentale focalizzato su un settore prefissato
- due possibilità
  - alimentato dal data warehouse primario
  - alimentato direttamente dalle sorgenti
- realizzazione più rapida
- richiede progettazione attenta, in modo da evitare problemi di integrazione in seguito

Server per data warehouse

- Server ROLAP (Relational OLAP)
  - DBMS relazionale esteso
    - rappresentazione compatta di dati sparsi
  - estensioni SQL per aggregati
  - metodi di accesso speciali che realizzano le operazioni di accesso in modo efficiente
- Server MOLAP (Multidimensional OLAP)
  - dati rappresentati in forma matriciale (multidimensionale) proprietaria
    - dati sparsi richiedono compressione
  - primitive OLAP speciali
- Server HOLAP (Hybrid OLAP)

Strumenti di back-end

- Alimentazione del data warehouse (ETL = Extraction Transformation Loading)
  - estrazione dei dati da sorgenti esterne
  - pulizia dei dati (errori, dati mancanti o duplicati)
  - trasformazioni e conversioni di formato
  - caricamento e refresh periodico

Metadati

Metadati = dati sui dati

- Diversi tipi di metadati
  - per trasformazione e caricamento: descrivono i dati sorgenti e le trasformazioni necessarie
    - utile usare una notazione comune per dati sorgente e dati risultanti dalle trasformazioni
    - CWMI (Common Warehouse Metadata Initiative): standard proposto da OMG per l'interscambio di dati tra strumenti DW e repository di metadati in ambienti eterogenei e distribuiti
  - per la gestione dei dati: descrivono la struttura dei dati presenti nel data warehouse
    - anche per dati derivati, quali le viste materializzate
  - per la gestione delle query: dati sulla struttura delle query e monitoraggio della loro esecuzione
    - codice SQL della query
    - piano di esecuzione
    - uso di memoria e CPU

Architetture per data warehouse

Elena Baralis
Politecnico di Torino

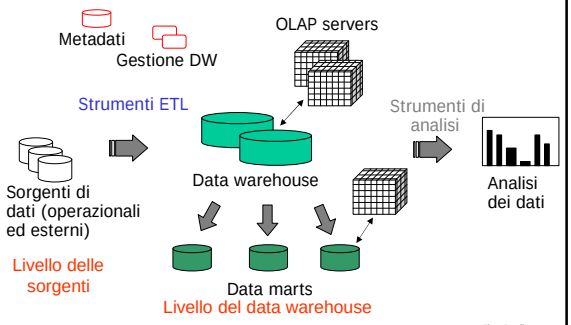
Architetture per data warehouse

- Separazione tra elaborazione transazionale e analisi dei dati
  - evitare le architetture a un livello
- Architetture a due o più livelli
  - separano in misura diversa i dati in ingresso nel data warehouse dai dati oggetto dell'analisi
  - maggiormente scalabili

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Architettura a due livelli



Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

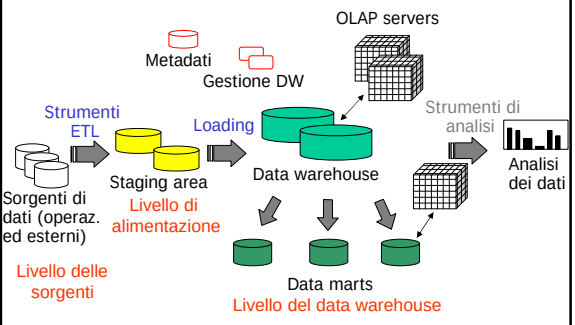
Caratteristiche delle architetture a 2 livelli

- Disaccoppiamento dalle sorgenti
  - possibilità di gestire dati esterni al sistema OLTP
  - modellazione dei dati adatta all'analisi OLAP
  - progettazione fisica del data warehouse mirata al carico analitico
- Facilità di gestione delle differenti granularità temporali dei dati operazionali e analitici
- Separazione del carico transazionale da quello analitico
- Necessità di svolgere "al volo" la preparazione dei dati (ETL)

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Architettura a tre livelli



Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Caratteristiche delle architetture a 3 livelli

- *Staging area:* area di transito che permette di separare l'elaborazione ET dal caricamento nel data warehouse
  - permette operazioni complesse di trasformazione e pulizia dei dati
  - offre un modello integrato dei dati aziendali, ancora vicino alla rappresentazione OLTP
  - talvolta denominata Operational Data Store (ODS)
- Introduce ulteriore ridondanza
  - aumenta lo spazio necessario per i dati

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Extraction, Transformation and Loading (ETL)

- Processo di preparazione dei dati da introdurre nel data warehouse
  - estrazione dei dati dalle sorgenti
  - pulitura
  - trasformazione
  - caricamento
- semplificato dalla presenza di una staging area
- eseguito durante
  - il primo popolamento del DW
  - l'aggiornamento periodico dei dati

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

ETL

Estrazione: acquisizione dei dati dalle sorgenti

Pulitura: operazioni volte al miglioramento della qualità dei dati (correttezza e consistenza)

Trasformazione: conversione dei dati dal formato operazionale a quello del data warehouse (integrazione)

Caricamento: propagazione degli aggiornamenti al data warehouse

Copyright – Tutti i diritti riservati

Struttura e elaborazione dei dati

Elena Baralis
Politecnico di Torino

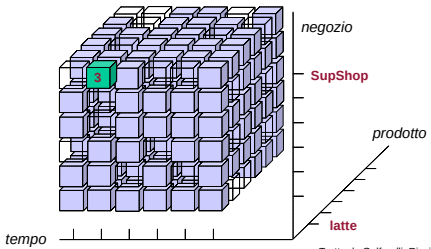
Copyright – Tutti i diritti riservati

Rappresentazione multidimensionale

- Dati rappresentati come un (iper)cubo con tre o più dimensioni
- Misure su cui si esegue l'analisi: elementi individuati all'intersezione delle dimensioni
- Data warehouse per l'analisi delle vendite di una catena di supermercati
  - assi dimensionali: prodotto, negozio, tempo
  - misure: quantità venduta, importo della vendita, …

Copyright – Tutti i diritti riservati

Rappresentazione multidimensionale



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

Copyright – Tutti i diritti riservati

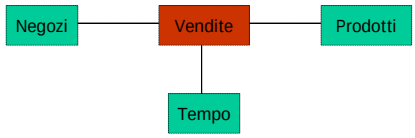
Rappresentazione relazionale: modello a stella

- Misure numeriche memorizzate nella *tabella dei fatti*
  - gli attributi contengono valori numerici
- Le *dimensioni* descrivono il contesto di ogni misura nella tabella dei fatti
  - contengono molti attributi descrittivi

Copyright – Tutti i diritti riservati

Esempio

Data warehouse per l'analisi delle vendite di una catena di supermercati



Copyright – Tutti i diritti riservati

Dimensione del data warehouse

- dimensione tempo: 2 anni x 365 giorni
- dimensione negozio: 300 negozi
- dimensione prodotto: 30.000 prodotti, di cui 3.000 venduti ogni giorno in ogni negozio
- numero di righe della tabella dei fatti:

  730 x 300 x 3000 = 657 milioni

⇨spazio occupato dalla tabella dei fatti ≈ 21GB

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Strumenti di analisi dei dati

- Analisi OLAP: calcolo di funzioni aggregate complesse
  - necessità di fornire supporto a diversi tipi di funzione aggregata (esempi: media mobile, top ten)
- Analisi dei dati mediante tecniche di data mining
  - varie tipologie di analisi
  - pesante componente algoritmica

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Strumenti di analisi dei dati

- Presentazione
  - attività distinta dalla ricerca: i dati ottenuti da una ricerca possono essere rappresentati mediante diversi tipi di strumenti di rappresentazione
- Ricerca di motivazioni
  - esplorazione dei dati mediante approfondimenti (esempio: drill down)

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Data mining: tipologie di attività

Classificazione e regressione: generazione di un modello predittivo
  - richiede un insieme di dati già classificati

Regole di associazione: ricerca di correlazioni tra i dati

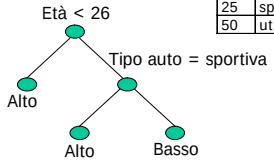
Clustering: suddivisione dei dati in gruppi “omogenei”
  - richiede una definizione di distanza

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Esempio: classificazione

Età	Tipo auto	Classe rischio
40	familiare	basso
65	sportiva	alto
20	utilitaria	alto
25	sportiva	alto
50	utilitaria	basso



Albero di decisione

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Esempio: regole di associazione

- E` data una collezione di transazioni di cassa di un supermercato (scontrini)
- Regola di associazione

  Pannolini ⇒ Birra
  - il 2% delle transazioni contiene entrambi gli elementi
  - il 30% delle transazioni che contengono pannolini contiene anche birra

Copyright – Tutti i diritti riservati

Elena Baralis
Politecnico di Torino

Libri

- Data warehousing
  - Golfarelli, Rizzi, *Data warehouse: teoria e pratica della progettazione*, McGraw-Hill 2006
  - Kimball e altri, *numerosi testi su metodologia e casi di studio,* Wiley
- Data mining
  - Han, Kamber, *Data mining: concepts and techniques*, Morgan Kaufmann 2006
  - Tan, Steinbach, Kumar, *Introduction to data mining*, Pearson 2006

Copyright – Tutti i diritti riservati

Link utili

- Data warehouse
  - http://www.dwinfocenter.org
  - http://www.dwreview.com
  - http://kimballuniversity.com
- Specifico per data mining
  - http://www.kdnuggets.com

Copyright – Tutti i diritti riservati