



Database and data mining group, Politecnico di Torino 

Data warehouse

Introduzione

Elena Baralis
Politecnico di Torino

Copyright – Tutti i diritti riservati INTRODUZIONE - 1 *Elena Baralis*
Politecnico di Torino

Database and data mining group, Politecnico di Torino 

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 INTRODUZIONE - 2 *Elena Baralis*
Politecnico di Torino

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

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

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

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 volatiliutilizzati per il supporto alle decisioni aziendali


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

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)

Struttura e analisi dei dati


Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino


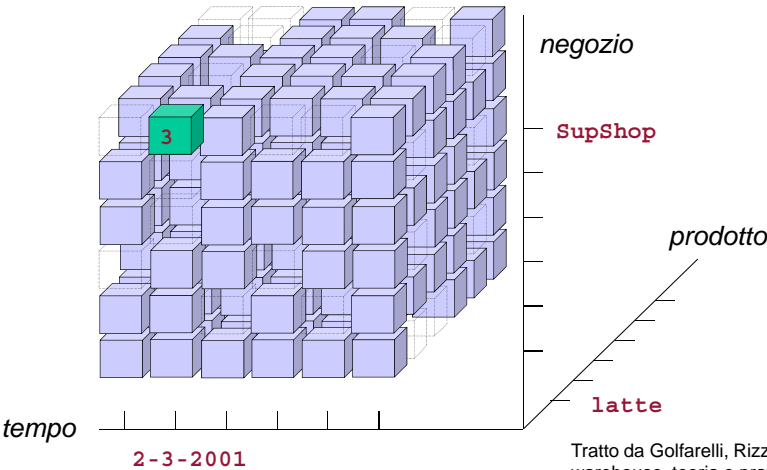
Rappresentazione multidimensionale

- 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
INTRODUZIONE - 9
Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino


Rappresentazione multidimensionale



Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006
 Elena Baralis
 Politecnico di Torino

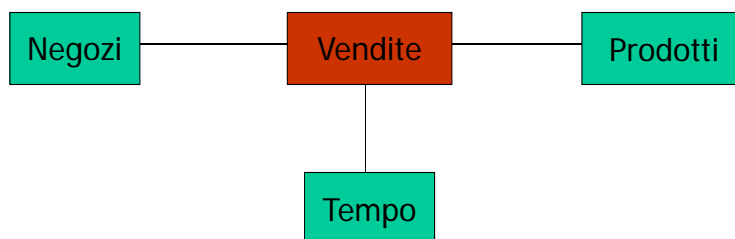
Copyright – Tutti i diritti riservati
INTRODUZIONE - 10

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

Esempio

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



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:
 $730 \times 300 \times 3000 = 657 \text{ milioni}$

⇒ spazio occupato dalla tabella dei fatti $\approx 21\text{GB}$

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

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)

Architetture per data warehouse

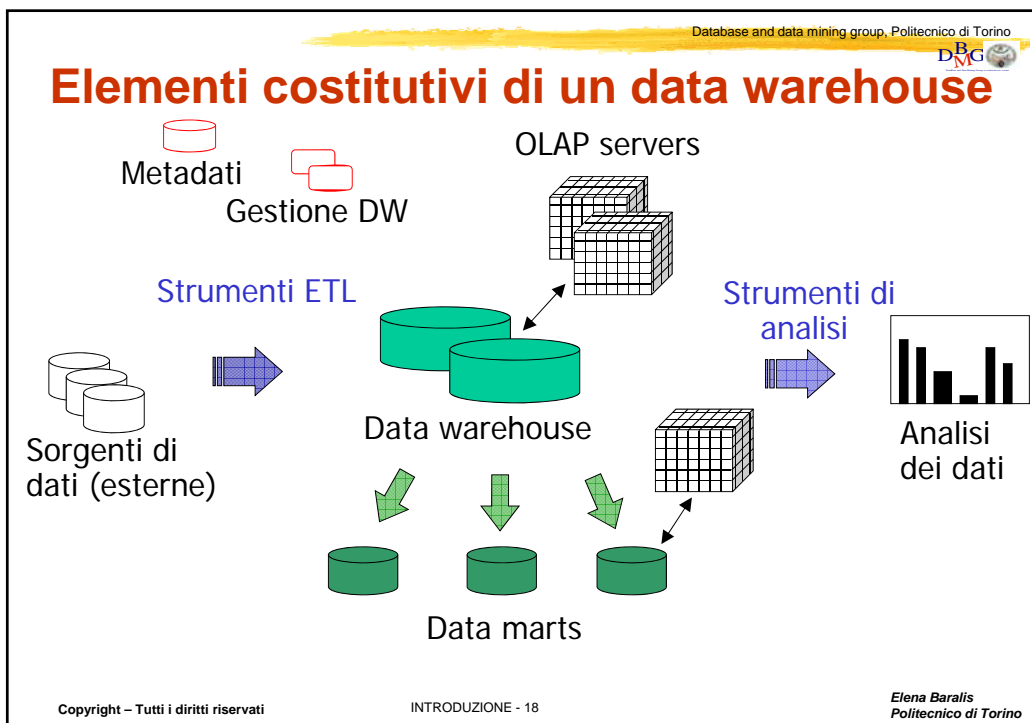
Elena Baralis
Politecnico di Torino

Database and data mining group, 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 INTRODUZIONE - 17 *Elena Baralis*
 Politecnico di Torino



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 ETL

- Processo di preparazione dei dati da introdurre nel data warehouse
 - Extraction
 - Transformation
 - Loading
- Processo eseguito durante
 - il primo popolamento del DW
 - l'aggiornamento periodico dei dati

Processo 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 operativo a quello del data warehouse (integrazione)

Caricamento: propagazione degli aggiornamenti al data warehouse

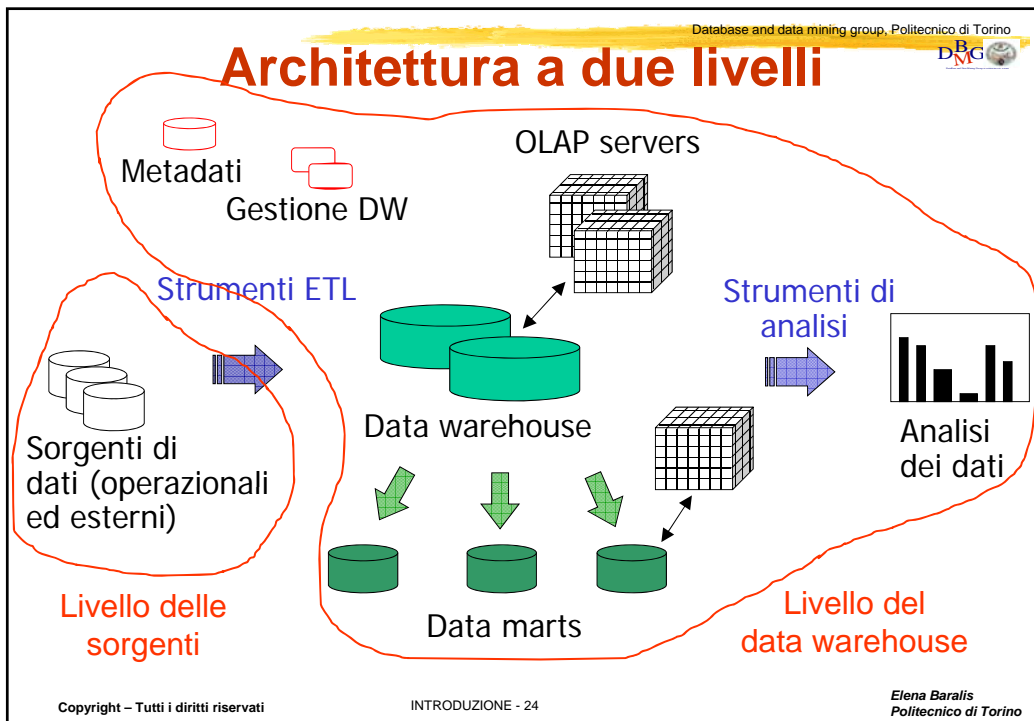
Database and data mining group, Politecnico di Torino


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

Copyright – Tutti i diritti riservati
INTRODUZIONE - 23
Elena Baralis
Politecnico di Torino

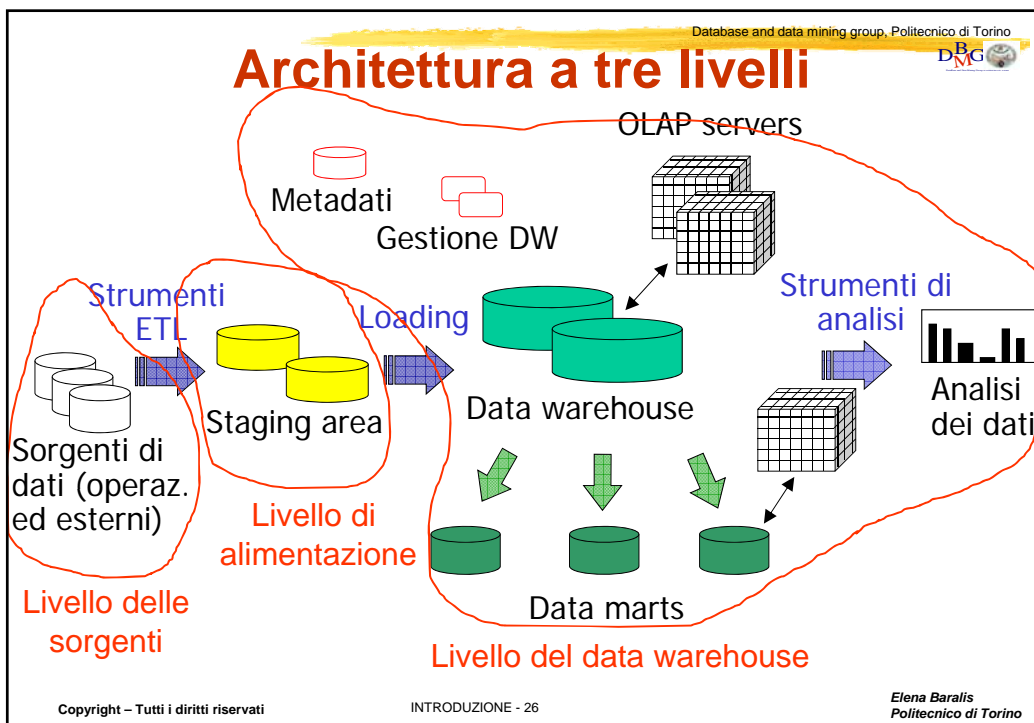


Database and data mining group, Politecnico di Torino
DBG

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 INTRODUZIONE - 25 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