Business Intelligence Technologies

Donato Malerba
Dipartimento di Informatica
Università degli Studi, Bari, Italy
malerba@di.uniba.it
http://www.di.uniba.it/~malerba

First page

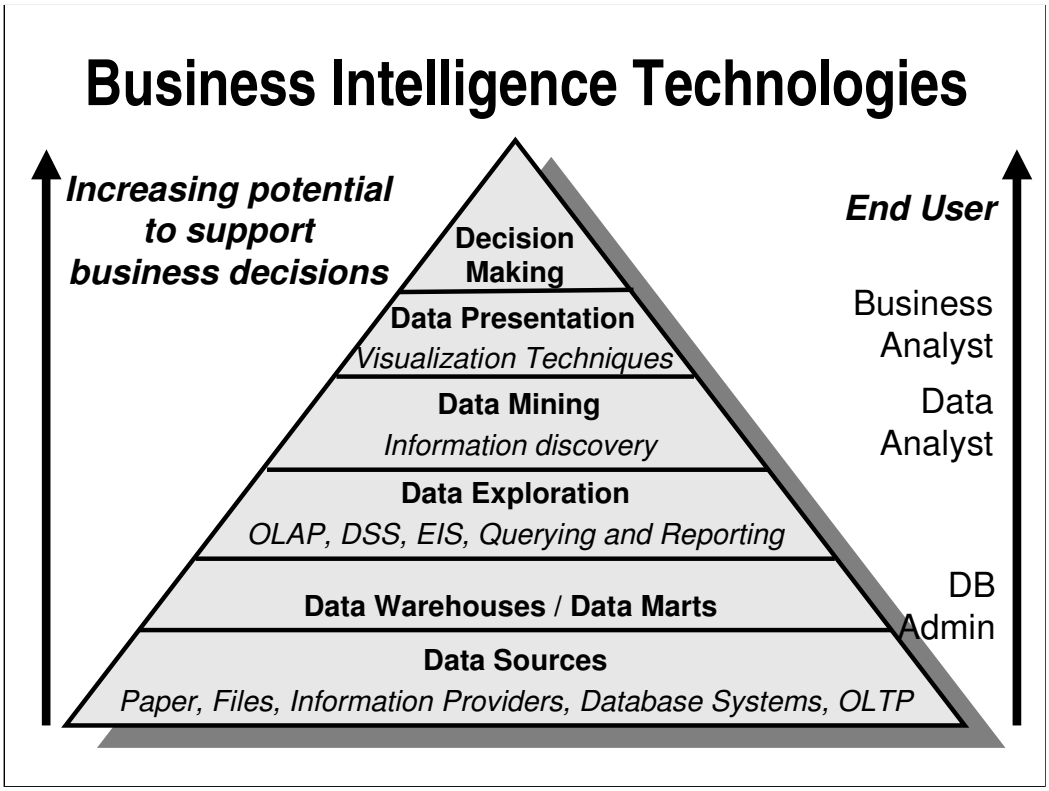
Business Intelligence

- Business Intelligence is a global term for all the processes, techniques and tools that support business decision-making based on information technology.
- The approaches can range from a simple spreadsheet to a major competitive undertaking.
- Data mining is an important new component of business undertaking.

First page

Un termine molto in voga negli ultimi tempi è quello di Business Intelligence, che si può definire come l'insieme dei processi, delle tecniche e degli strumenti basati sulla tecnologia dell'informazione e che supportano i processi decisionali di carattere economico. Il problema che sottostà al Business Intelligence è quello di avere sufficienti informazioni in modo tempestivo e fruibile e di analizzarle cosicché da poter avere un impatto positivo sulle strategie, le tattiche e le operazioni di business. Le informazioni riguardano la specifica impresa oppure situazioni più generali di mercato. Quando le informazioni riguardano esclusivamente la concorrenza si parla anche di Competitive Intelligence. Un altro termine ampiamente adottato è quello di Knowledge Management, che riguarda come le organizzazioni creano, catturano e riusano la conoscenza per raggiungere pero' obiettivi organizzativi (e non strategici).

Tanto nel Business Intelligence quanto nel Knowledge Management è possibile individuare tre attività importanti: raccolta, analisi e proposta di consigli. Gli strumenti utilizzati per queste attività possono andare dal semplice foglio elettronico a sistemi per il data mining.



Questa figura illustra la disposizione logica delle differenti tecnologie adottate nel business intelligence. La disposizione si basa sul potenziale valore delle tecnologie come base per decisioni strategiche e di business.

In generale il valore dell'informazione a supporto dei processi decisionali cresce dal basso verso l'alto. Una decisione basata sui dati nei livelli più bassi, dove tipicamente ci sono milioni di record, influenzerà la transazione di un singolo cliente, mentre una decisione presa su dati fortemente riepilogati come quelli dei livelli più alti riguarderà probabilmente un intero dipartimento o persino l'intera azienda.

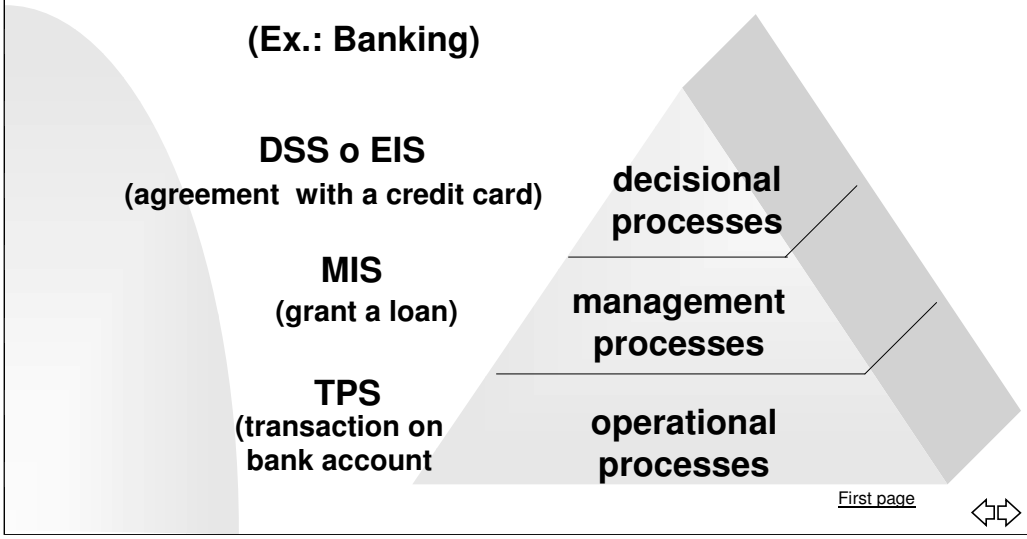
Per questa ragione si trovano generalmente differenti tipi di utenti sui diversi livelli. Un amministratore di basi di dati lavora principalmente con i database sulla sorgente dei dati e sul livello di data warehouse, mentre l'analista economico o la dirigenza lavorerà principalmente sui livelli più elevati della piramide.

Va ricordato che questo è una disposizione logica e non una interdipendenza fisica tra i vari livelli tecnologici. Per esempio le tecniche di visualizzazione possono essere usate indipendentemente dal data mining, il data mining può basarsi sui data warehouse o semplicemente su file. I data warehouse sono una tecnologia di supporto al data mining e non una condizione essenziale. Infatti molte delle applicazioni di data mining oggi sono effettuate comunque su file estratti direttamente da sorgenti operazionali dei dati. La connessione fra data warehousing e data mining è comunque piuttosto forte, ragione per cui essa verrà approfondita in questa sede.

Business Processes

- Data for support decision making
- Different information systems support the different processes

(Ex.: Banking)



Iniziamo col dire che i data warehouse sono stati pensati per fornire i dati di supporto ai processi decisionali.

Infatti le attività svolte in una organizzazione possono essere raggruppate come segue:

• gestione dei rapporti con l'esterno, dovuti a servizi offerti o a prodotti scambiati (processi produttivi o operativi). Ad esempio, in una banca un processo operativo è la transazione su un conto corrente.

• gestione operativa dell'organizzazione, ovvero gestione e controllo delle risorse e delle loro modalità di utilizzo (processi gestionali). In una banca, un processo gestionale è la concessione di un mutuo. La risorsa, in questo caso, è la disponibilità finanziaria dell'istituto di credito.

• attività di programmazione per fissare priorità di interventi (processi decisionali o di governo). Sempre con riferimento alla banca, un processo decisionale è la stipula di un accordo con una carta di credito.

E' quest'ultimo tipo di processo che il data warehouse intende supportare.

La diversità dei dati operazionali da quelli decisionali, detti anche business data, ha portato allo sviluppo di diversi sistemi informatici di supporto agli specifici processi di un'organizzazione.

I Decision Support System e gli Executive Information System sono utilizzati nei processi decisionali di più alto livello, mentre i Management Information System sono utilizzati sono settoriali e sono utilizzati per i processi gestionali, e infine i Transaction Processing Systems sono utilizzati nei processi operativi.

DSS vs. EIS

- Decision Support Systems (DSS) and Executive Information Systems (EIS): information systems designed to help managers in making choices.
- Different, yet interrelated applications
- A DSS focuses on a particular decision, whereas an EIS provides a much wider range of information (e.g., information on financials, on production history, and on external events).
- DSSs appeared in the 1970s
- EISs appeared in the 1980s.

First page

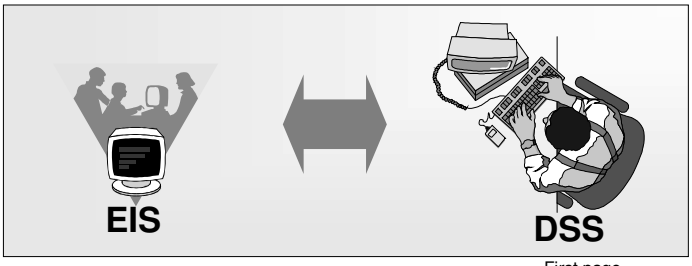
DSS e EIS sono applicazioni differenti ma allo stesso tempo correlate.

Un DSS vennero sviluppati negli anni ‘70. Questi usavano modelli matematici anche complessi per *risolvere* problemi manageriali. L’idea era che la direzione di un’organizzazione dovesse utilizzare questi sistemi in modo autonomo. Questa assunzione si dimostrò subito falsa, poiché molti dei manager non avevano gli skill e il tempo per utilizzare questi sistemi. Pertanto a partire dagli inizi degli anni ‘80, molte organizzazioni e vendor offrirono dei sistemi informativi più semplici, chiamati Executive Information System.

L’idea alla base degli EIS è che la direzione è interessata ad avere informazioni standard sulla propria azienda e sull’ambiente esterno. Così un EIS includeva informazioni sul bilancio, sulla storia della produzione, sul personale, come pure notizie sui concorrenti.

DSS vs. EIS

- The original EISs did not have the analytical capabilities of a DSS
- “An EIS is used by senior managers to find problems; the DSS is used by the staff people to study them and to offer alternatives” (Rockart and Delong, 1988)



First page

I primi EIS non avevano alcuna capacità di analisi dei dati, a differenza dei DSS. Come puntualizzato da Rockart e Delong alla fine degli anni ‘80, un EIS è utilizzato da manager anziani per trovare i problemi, mentre il DSS è utilizzato dallo staff per offrire delle alternative.

La differenza di utenti comporta evidentemente una differenza anche del tipo di interfaccia offerta dal sistema.

Where do Data Come From?

- The EISs and DSSs often lacked a strong database component.
- Most organizational information gathering was (and is) directed to maintaining current (preferably on-line) information about individual transactions and customers.
- Managerial decision making requires consideration of the past and the future, not just the present.
- New databases, called *data warehouses*, were created specifically for analytic use

First page

Inizialmente, tanto gli EIS quanto i DSS erano carenti di una componente atta a raccogliere in maniera integrata e permanente i dati di interesse. Insomma, mancava un database.

La ragione era che i dati raccolti tipicamente all'interno di una organizzazione erano (e sono) orientati a supportare decisioni in processi operativi, o al più, in processi gestionali.

Tali dati fotografavano la situazione corrente di un'organizzazione, mentre alla dirigenza occorrevano anche dati relativi al passato.

Le case produttrici di database si resero conto negli anni '80 che i loro principali sforzi erano stati compiuti per sviluppare database adatti a supportare elaborazioni transazionali, mentre i database a supporto dei processi decisionali di alto livello dovevano supportare elaborazioni di tipo analitico.

Si cominciò pertanto a sviluppare nuovi database, chiamati data warehouse, cioè magazzini di dati, perché dovevano servire a contenere grandi quantità di dati, ben più di quelle necessarie per elaborazioni transazionali, per un lungo periodo di tempo.

A Data Warehouse is ...

A data warehouse is a

- **subject-oriented,**
- **integrated,**
- **time-variant, and**
- **nonvolatile**

collection of data in support of management's decisions

Inmon, W.H.
Building the Data Warehouse
Wellesley, MA: QED Tech. Pub. Group, 1992

First page

La definizione di Inmon è quella più diffusamente riconosciuta e focalizza l'attenzione su una serie di aspetti peculiari del data warehouse,

Il data warehouse, come collezione di dati a supporto del processo decisionale del management, è:

• orientato al soggetto

• integrato

• invariante nel tempo

• non volatile

… subject-oriented ...

- The data in the warehouse is defined and organized in business terms, and is grouped under *business-oriented subject headings*, such as
  - **customers**
  - **products**
  - **sales**

  rather than individual transactions.
- *Normalization* is not relevant.

First page

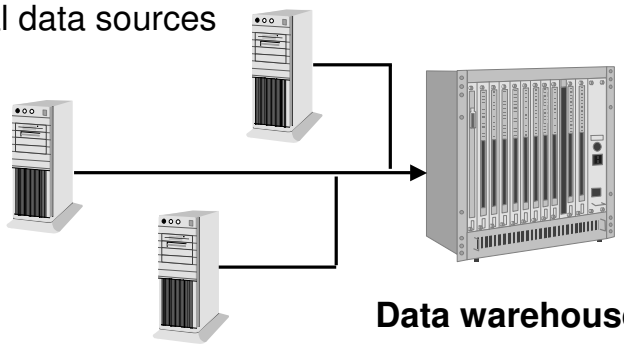
Orientato al soggetto significa che i dati vanno raggruppati per aree di interesse come i clienti, i fornitori, i prodotti piuttosto che per processi operativi o applicazioni come marketing e vendita. I dati di un data warehouse sono finalizzati a chi li usa più che a chi li genera, a differenza delle tradizionali basi di dati il cui disegno è orientato ai requisiti elaborativi dei singoli processi operativi o applicazioni.

La normalizzazione dello schema dei dati, ad esempio, è importante nelle basi di dati progettate per scopi transazionali, poiché la conseguente eliminazione delle ridondanze consente ad una generica transazione di poter aggiornare le singole informazioni di interesse in una sola parte del database.

Poiché la normalizzazione determina anche frammentazione dello schema dei dati, appare controproducente nella progettazione di un sistema di data warehousing.

… integrated …

- The data warehouse contents are defined such that they are valid across the enterprise and its operational and external data sources



Data warehouse

Operational systems

- The data in the warehouse should be
  - **clean**
  - **validated**
  - **properly integrated**

First page

Il data warehouse dev'essere integrato, ossia consistente rispetto ad uno schema concettuale globale dei dati,che abbraccia l'intera impresa e non i singoli dipartimenti.

Infatti un'assunzione implicita della definizione di Inmon è che il data warehouse è fisicamente separato dai database operazionali e da questi viene comunque alimentato. La separazione è giustificata dal fatto che

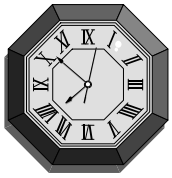
• non esiste un'unica base di dati operazionale che contiene tutti i dati di interesse

• non è tecnicamente possibile fare l'integrazione in linea

• i dati di interesse sarebbero comunque differenti

• degrado generale delle prestazioni senza la separazione.

… time-variant ...



- All data in the data warehouse is time-stamped at time of entry into the warehouse or when it is summarized within the warehouse.
- This chronological recording of data provides historical and trend analysis possibilities.
- On the contrary, operational data is overwritten, since past values are not of interests.

First page

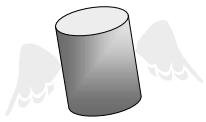
Le basi di dati operazionali mantengono il valore corrente di un dato e l'orizzonte temporale di interesse è dell'ordine di due-tre mesi.

Al contrario nel data warehouse è di interesse l'evoluzione storica dei dati, con un'orizzonte temporale dell'ordine di anni. Disponendo di informazioni sul passato è possibile, attraverso gli strumenti di supporto decisionale e di data mining, capire le tendenze future (analisi del trend).

Per poter rappresentare l'evoluzione del dato, si applica un timestamp (orario di stampa) ai dati operazionali, a livello di singoli campi o di interi record.

Ovviamente questo significa che le chiavi delle entità memorizzate in un data warehouse conterranno un elemento di tempo.

… nonvolatile ...



- Once loaded into the data warehouse, the data is not updated.
- Data acts as a stable resource for consistent reporting and comparative analysis.
- On the contrary, operational data is updated (inserted, deleted, modified).

First page

Il dato caricato in un data warehouse può essere ispezionato ma non modificato dall'utente.

Nel data warehouse si hanno operazioni di accesso e interrogazione effettuate di giorno dagli utenti, ed operazioni di caricamento e aggiornamento dei dati effettuate automaticamente in orari notturni.

Which Data in the Warehouse?

- A data warehouse contains five types of data:
  - **Current detail data**
  - **Old detail data**
  - **Lightly summarized data**
  - **Highly summarized data**
  - **Metadata**
- *Granularity* of the data: a key design issue

First page

Un data warehouse contiene cinque tipi di dati.

I dati dettagliati attuali. Il livello di dettaglio potrebbe non essere lo stesso delle basi di dati operazionali, poiché il volume di questi potrebbe essere anche eccessivo. Inoltre, non tutti i dati disponibili nelle basi di dati operazionali vanno riportati nel warehouse. Inoltre i dati attuali potrebbero non essere i più recenti, dato che il caricamento dei dati nel warehouse non avviene on-line.

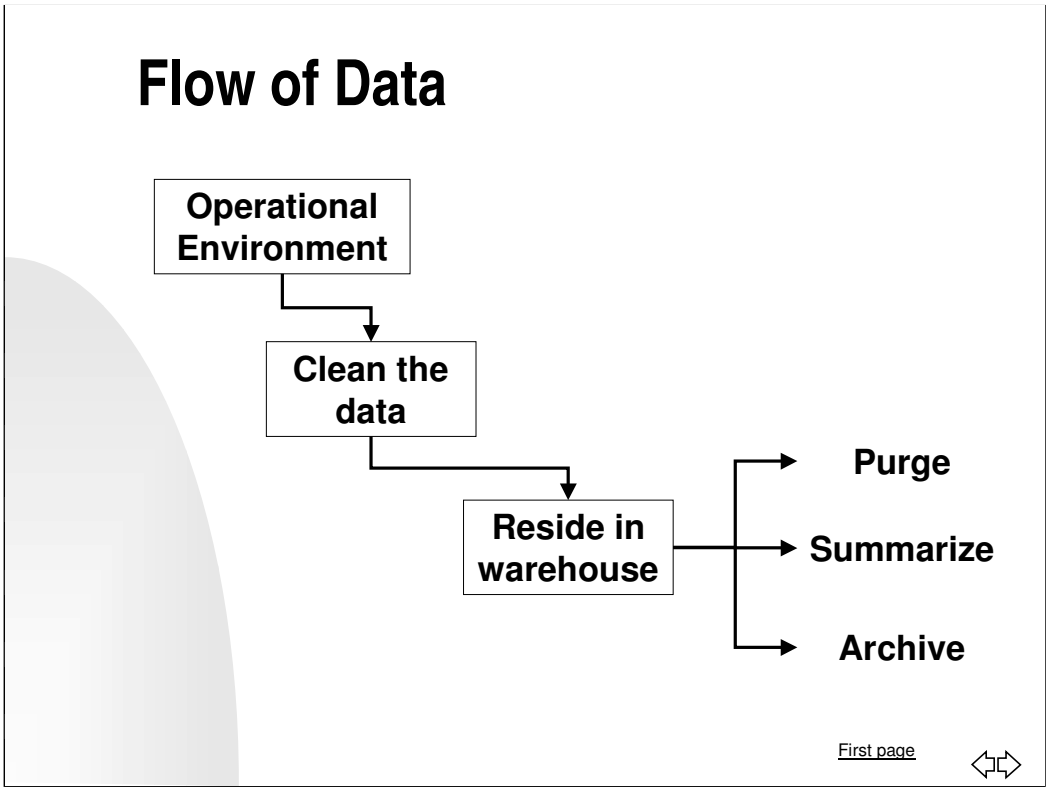
Evidentemente, a causa delle dimensioni in gioco, non tutte le informazioni possono essere tenute in linea, e tipicamente i dati storici dettagliati saranno archiviati off-line: quindi le risposte ad interrogazioni che riguardano i livelli di sintesi più alti saranno senz'altro interattive, mentre analisi che si spingono fino al reperimento di informazioni sui dati storici archiviati non potranno che scatenare processi batchm determinando risposte differite.

Molte applicazioni di supporto alle decisioni operano su dati sintetici ottenuti per riepilogo dei dati transazionali. Riepilogando i dati in anticipo si migliorano i tempi di risposta. Il progettista di data warehouse deve decidere:

• quali attributi riepilogare

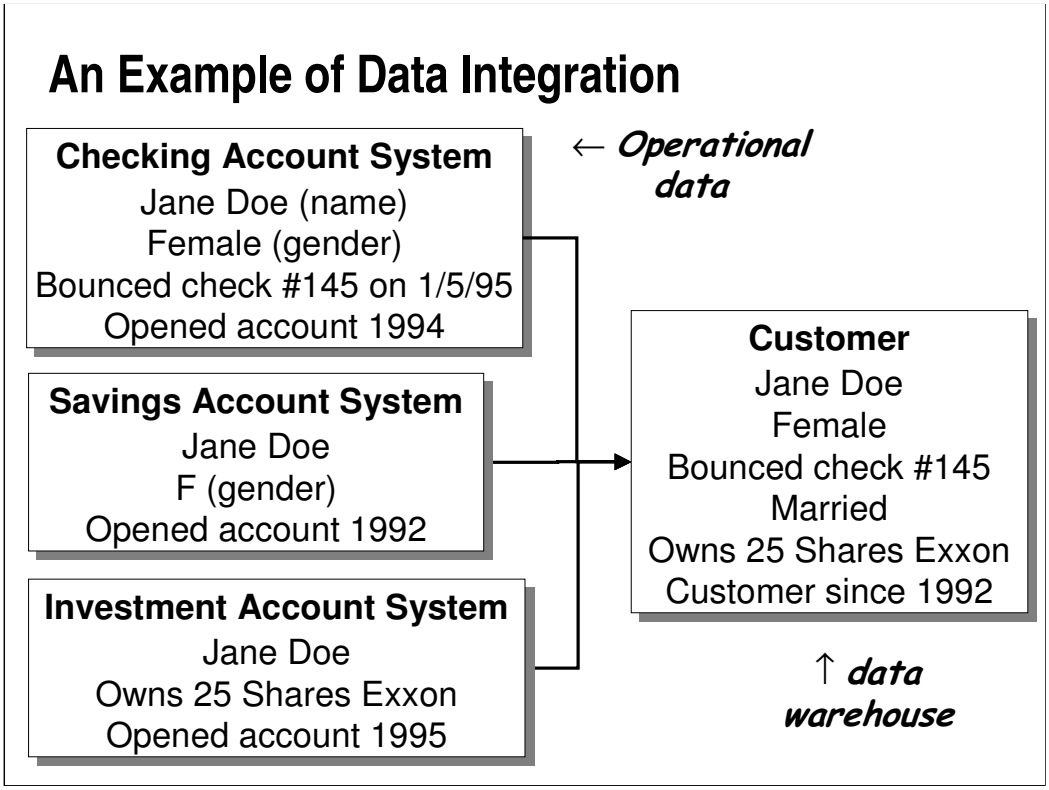
• quale unità di tempo scegliere per il riepilogo.

I metadati definiscono il contenuto del warehouse, specificano le regole usate per il riepilogo dei dati e offrono una guida al mapping dalla forma operazionale alla forma del warehouse.



Quasi tutti i dati entrano nel warehouse dall'ambiente operazionale. I dati sono quindi puliti e spostati nel warehouse. I dati continuano a risiedere nel warehouse finche non è intrapresa una delle tre azioni:

- I dati sono rimossi
- I dati sono riepilogati
- I dati sono archiviati.



Quasi tutti i dati entrano nel warehouse dall'ambiente operazionale. I dati sono quindi puliti e spostati nel warehouse. I dati continuano a risiedere nel warehouse finche non è intrapresa una delle tre azioni:

- I dati sono rimossi
- I dati sono riepilogati
- I dati sono archiviati.

Cost and Size of a Data Warehouse

- Data warehouses are expensive undertakings (*mean cost: $2.2 million*).
- Since a data warehouse is designed for the enterprise it has a typical storage size running *from 50 Gb to over a Terabite*.
- *Parallel computing* to speed up data retrieval

WAREHOUSE SIZE	*SERVER REQUIREMENTS*
5-50 GB	Pentium PC > 100MHz
50-500 GB	SMP machine
> 500 GB	SMP or MPP machine

First page

La realizzazione di un data warehouse è un'impresa costosa, dell'ordine dei 2,2 milioni di dollari in media (stima del 1996).

Il costo è dovuto non solo alla progettazione e manutenzione ma anche alla piattaforma hardware/software, che deve consentire la memorizzazione dai 50Gb fino al Terabite di dati. Al crescere della dimensione del data warehouse diventa indispensabile ricorrere al calcolo parallelo con processori a memoria condivisa e non.

SMP: shared memory multiprocessor

MPP: massively parallel multiprocessors

Raccomandazion del Meta Group. DePompa, B. (1996) "Stack that Data," *Information Week*. January 29, 1996, pp. 50-57.

The Data Mart

- A lower-cost, scaled-down version of the data warehouse designed for the strategic business unit (SBU) or department level.
- An excellent first step for many organizations.
- Main problem: data marts often differ from department to department.
- Two approaches:
  - **data marts ⇒ enterprise-wide system**
  - **data warehouse ⇒ data marts**

First page

L'elevato costo dei data warehouse ne limita il loro uso alle grandi organizzazioni.

Un data mart è una versione in scala di un data warehouse, progettato per uno specifico settore o dipartimento. Ad esempio, un data mart potrebbe contenere solo i dati relativi al marketing.

I data mart presentano diversi vantaggi rispetto ai data warehouse

•Il costo è basso (meno di 100 mila dollari)

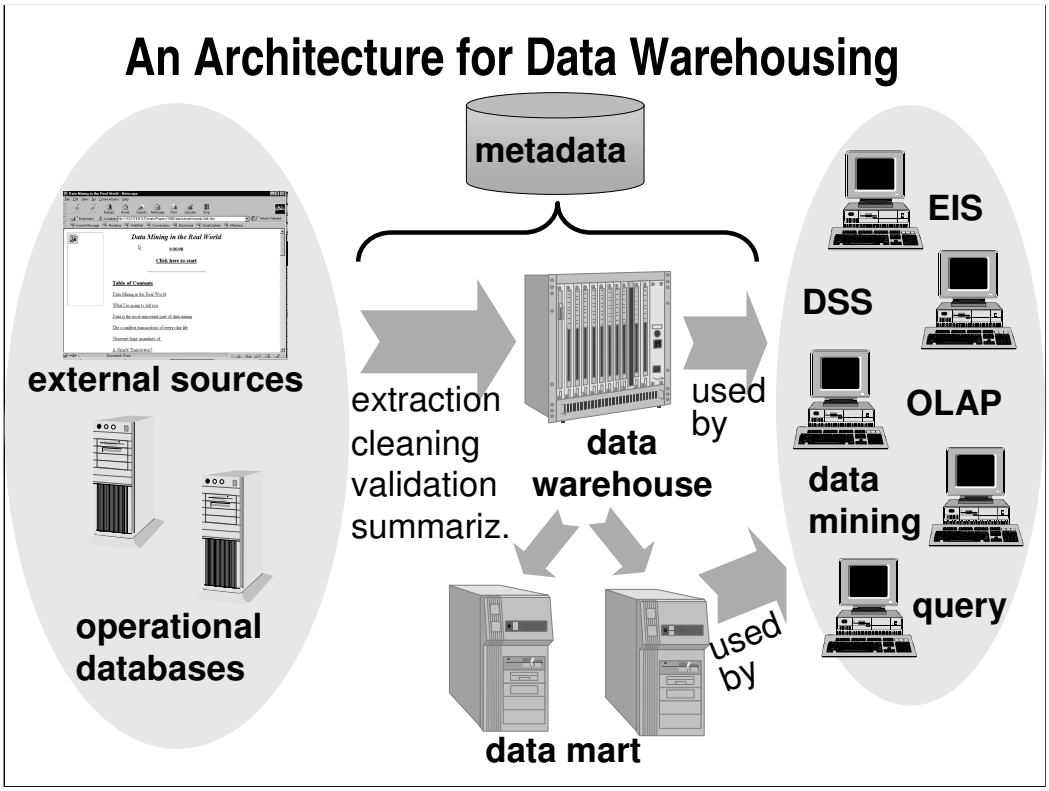
•Il tempo di implementazione è ridotto (pochi mesi)

• Sono controllati localmente piuttosto che centralmente, conferendo potere al gruppo che lo utilizza.

• Sono più piccoli e dunque con tempi di risposta più rapidi.

Una idea di data mart è il prodotto IBM Visual Warehouse 1.2, con una capacità dai 10 ai 50 GB.

Il problema dei data mart è che essi possono differire notevolmente. Per cui lo sviluppo di un data warehouse a livello di impresa può richiedere una fase di integrazione. Questo se si sviluppa il data warehouse dopo aver progettato dei data mart. Un approccio alternativo è quello di progettare prima un warehouse centralizzato e poi in base a questo progettare dei data mart settoriali, con prestazioni migliori.



L'architettura per il data warehousing prevede innanzitutto delle sorgenti dei dati, come le basi di dati dei sistemi informativi operazionali, detti anche *legacy system*, ed eventualmente delle sorgenti esterne all'azienda (ad esempio il Web).

Il caricamento del data warehouse comprende sia quello iniziale (*initial load*) e gli aggiornamenti periodici (*trickle feed*). Le operazioni sono di estrazione, con accesso ai dati esterni, di trasformazione con pulizia dei dati, di validazione e di riepilogo.

I metadati sono informazioni mantenuti a supporto di queste attività.

I dati caricati nel warehouse possono essere poi utilizzati per alimentare i data mart.

Dati del warehouse e dei data mart sono utilizzati dagli strumenti di analisi che possono andare da semplici linguaggi di interrogazione, a DSS, EIS, strumenti di interrogazione e OLAP.

Architettura di una DW

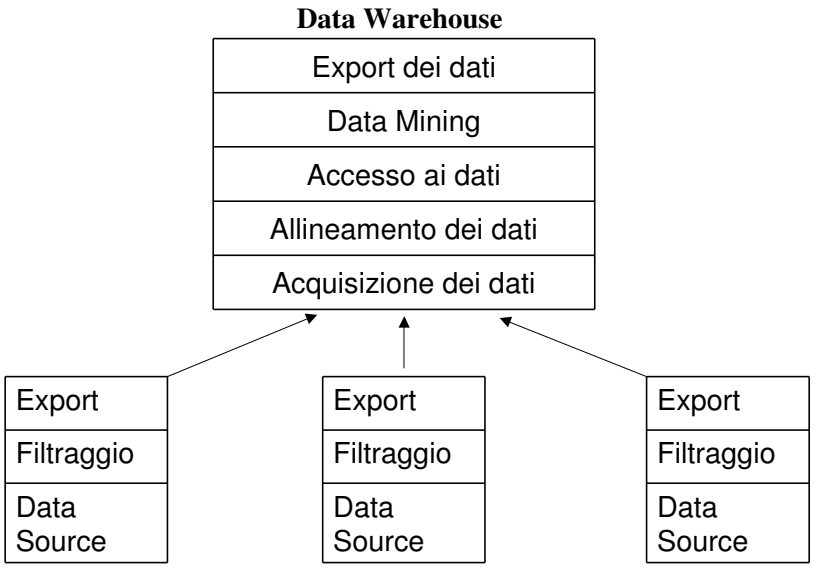
Una DW contiene dati che vengono estratti da uno o più sistemi, detti *data source*, incluso:

- raccolte dati non gestite tramite DBMS
- raccolte dati gestiti da DBMS di vecchia concezione (*legacy system*)

Nell'architettura di una DW si possono identificare dei componenti nella data warehouse e altri esterni, nelle *data source*. Questi sono:

1. Un componente di filtraggio dei dati
2. Un componente per l'esportazione dei dati

Architettura di una DW



Componente di filtraggio dei dati

- I filtri controllano la correttezza dei dati prima dell'inserimento nella data warehouse.
- I filtri possono *eliminare* dati palesemente scorretti sulla base di vincoli e controlli che si applicano a singole data source
- Talvolta possono *rilevare* e *correggere* inconsistenze nei dati estratti da molteplici data source.
- In tal modo viene fatta la pulizia dei dati (*data cleaning*) che è essenziale per assicurare un sufficiente livello di qualità.

Componente per l'esportazione dei dati

- Consente di estrarre i dati dalla data source.
- In genere, il processo di esportazione è *incrementale*: il sistema per l'esportazione dei dati colleziona le sole modifiche (inserzioni o cancellazioni) delle data source, che vengono poi importate dalla DW.

Componente di acquisizione dei dati (loader)

I successivi cinque componenti operano nella data warehouse.

Il componente di acquisizione dei dati (*loader*) è responsabile di caricare inizialmente i dati nella DW.

Che fa?

Predispone i dati all'uso operativo, svolgendo nel contempo operazioni di ordinamento e aggregazione e costruendo le strutture dati della DW.

Se la DW ha un'architettura distribuita questo modulo si occupa anche della frammentazione iniziale dei dati.

Componente di acquisizione dei dati (loader)

Quando opera?

Tipicamente le operazioni di acquisizione vengono svolte a lotti (“batch”) quando la DW non è utilizzata per l’analisi (in genere di notte).

Se il volume dei dati è piccolo, questo modulo è invocato periodicamente per acquisire l’intero contenuto della DW.

Più spesso, dopo il caricamento iniziale, i dati vengono allineati in modo incrementale, utilizzando il modulo di allineamento dei dati.

Componente di allineamento dei dati (refresh)

Propaga incrementalmente le modifiche della *data source* in modo da aggiornare il contenuto della DW.

Si possono usare due tecniche:

- **invio dei dati** (***data shipping***): utilizza i trigger collocati nella data source che, in modo trasparente alle applicazioni, registrano inserimenti, cancellazioni, e modifiche in opportuni *archivi variazionali*; le modifiche vengono trattate come una coppia di inserimenti e cancellazioni.
- **invio delle transazioni** (***transaction shipping***): utilizza i log di transazione per costruire archivi variazionali.

Componente di allineamento dei dati (refresh)

In entrambi i casi gli archivi transazionali vengono poi utilizzati per *rinfrescare* (*refresh*) la DW, aggiungendo i dati nell'archivio variazionale degli inserimenti.

Per quanto concerne le cancellazioni, in genere i dati corrispondenti della DW vengono marcati come dati storici ma non cancellati.

First page

Componente per l'accesso ai dati

È responsabile di realizzare le operazioni necessarie all'analisi dei dati.

Nella DW, questo modulo realizza in modo efficiente interrogazioni complesse, caratterizzate da join tra tabelle, ordinamenti e aggregazioni complesse.

Esso consente anche nuove operazioni, quali roll-up e drill-down, per supportare funzionalità tipiche dell'OLAP.

First page

Componenti di data mining e di esportazione dei dati

- Il componente di *data mining*, che consente di svolgere ricerche sofisticate sulle informazioni “nascoste” nei dati.
- Il componente per l’*esportazione dei dati*, che consente di esportare i dati presenti in una warehouse ad altre DW, realizzando così un’architettura gerarchica.

First page

Moduli di ausilio

- Un componente per l'*assistenza allo sviluppo* della datawarehouse, che consente di definire lo schema dei dati e i meccanismi per l'importazione dei dati (ausilio alla progettazione).
- Un *dizionario dei dati*, che descrive il contenuto della DW, utile per comprendere quali analisi dei dati possono essere eseguite.

First page

La qualità dei dati

- La qualità dei dati è un elemento essenziale per il successo di una DW.
  - **Se i dati memorizzati contengono imprecisioni o errori, l'analisi risultante sarà necessariamente fuorviante, e l'uso della DW potrà risultare addirittura controproducente.**
- I fattori che pregiudicano la qualità dei dati sono:
  - **In basi di dati prive di vincoli di integrità, ad esempio perché gestite con tecnologie pre-relazionali, il tasso di errori (*dirty data*) è assai elevato (5÷30%)**
  - **In DW costruite assemblando dati estratti da plurime fonti si aggiungono problemi di disallineamento dei dati.**
- L'impiego di filtri e l'osservazione del processo di produzione dati sono fondamentali.

Modello Multidimensionale

- I dati di una DW sono organizzati per *aree di interesse* (dimensioni di analisi).
- Il modello ***multidimensionale*** sembra essere quindi il più consono per la modellazione *concettuale* di un DW. In un modello multidimensionale i dati sono organizzati in uno o più ***cubi multidimensionali*** (*data cube*).
- A differenza dei classici array, nei quali gli indici sono caratterizzati da un ordine lineare (tipicamente gli indici sono valori interi), in un modello multidimensionale sugli indici potrebbe non essere definito un ordine, o comunque si potrebbe definire un ordinamento parziale (gerarchie di valori).

Modello Multidimensionale

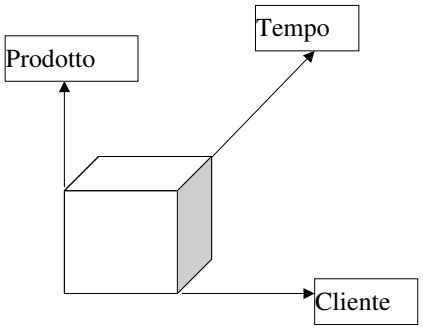
- Un cubo multidimensionale è incentrato su un fatto di interesse per il processo decisionale.
- Esso rappresenta un insieme di eventi, descritti quantitativamente da ***misure numeriche***.
- Ogni asse del cubo rappresenta una possibile dimensione di analisi; ciascuna dimensione può essere vista a più livelli di dettaglio individuati da attributi strutturati in gerarchie.

Modello Multidimensionale

Esempio: vendite di alcuni prodotti.

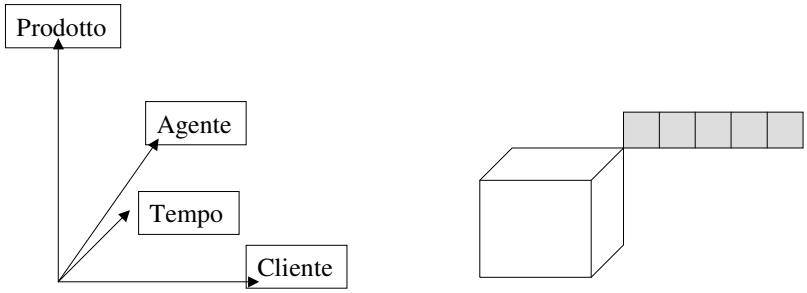
Le dimensioni in base alle quali le vendite vengono analizzate possono essere:

- i prodotti
- il tempo
- i clienti



Modello Multidimensionale

In realtà, le dimensioni di analisi potrebbero essere più di tre. Ad esempio, le vendite potrebbero essere analizzate considerando anche gli agenti che hanno intrapreso la trattativa. In questi casi si viene a creare un *ipercubo*.



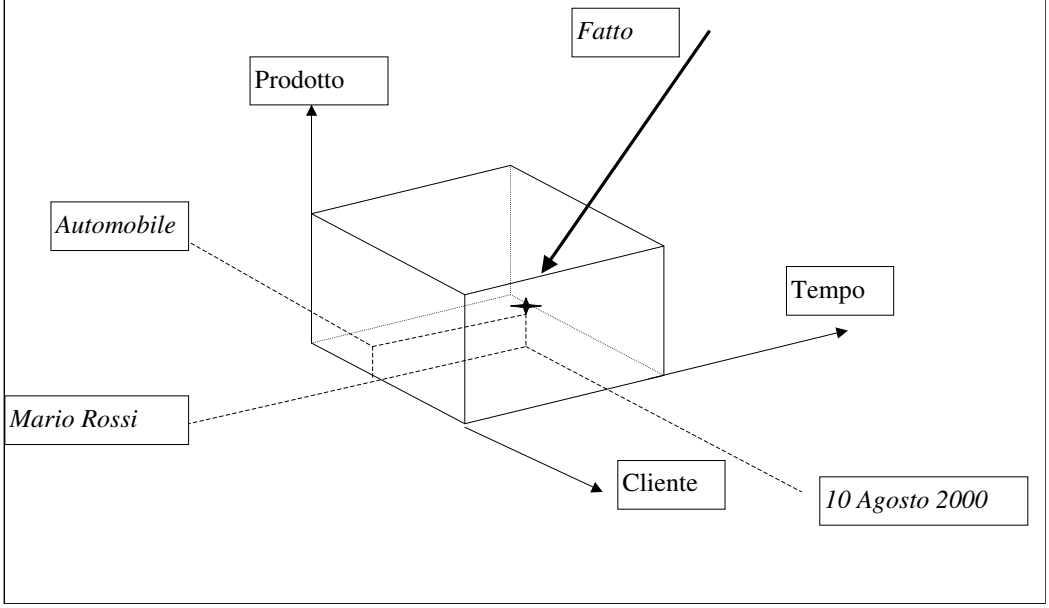
Modello Multidimensionale

Per accedere ai dati di una vendita è necessario specificarne le coordinate, ovvero dei valori per le dimensioni di analisi.

Esempio: referenziare la vendita del *10 agosto 2000*, del prodotto *automobile*, del cliente *Mario Rossi* o ancora selezionare tutte le vendite del mese di Agosto. In questo modo è possibile selezionare dall'ipercubo solo una porzione dei dati in esso esistenti.

Se per ciascuna delle dimensioni viene specificato un valore ben preciso, allora nell'ipercubo verrà individuata un'unica ***cella*** o un singolo ***fatto*** (nel nostro caso una vendita).

Modello Multidimensionale



Modello Multidimensionale

Se si fissa un valore ben preciso solo per una delle dimensioni si determina una ***fetta*** (***slice***) dell'ipercubo.

Le operazioni di interrogazione in un modello logico multidimensionale si riducono, quindi, a semplici selezioni di porzioni di un ipercubo.

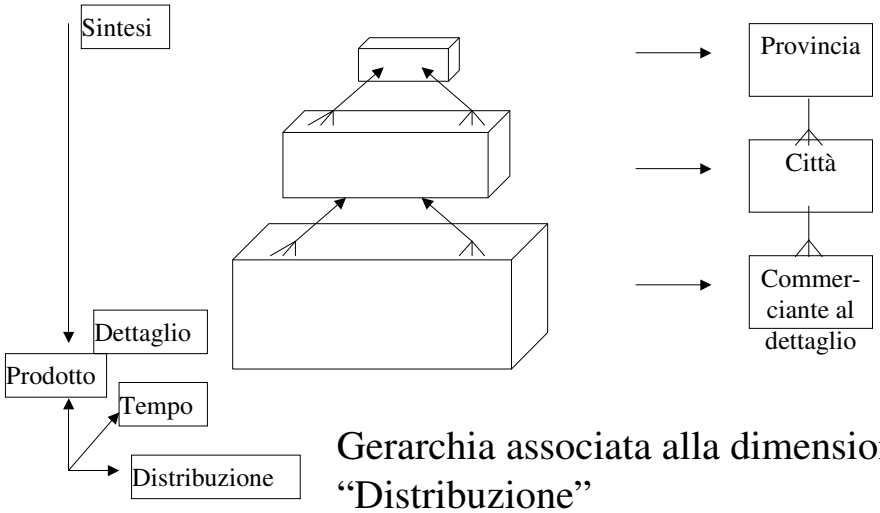
In tal senso, esse risultano più semplici delle interrogazioni di basi di dati relazionali, per le quali occorre spesso ricorrere a complesse operazioni di ***giunzione*** (***join***) con costi computazionali non trascurabili.

Modello Multidimensionale

Nel modello multidimensionale le dimensioni di analisi possono generalmente essere organizzate in **gerarchie**.

Esempio: la dimensione di analisi “distribuzione”, che fa riferimento alla rete di distribuzione di una certa azienda, avrà al livello più basso i commercianti al dettaglio, e ai livelli superiori le città e le province raggiunte dalla rete di distribuzione.

Modello Multidimensionale



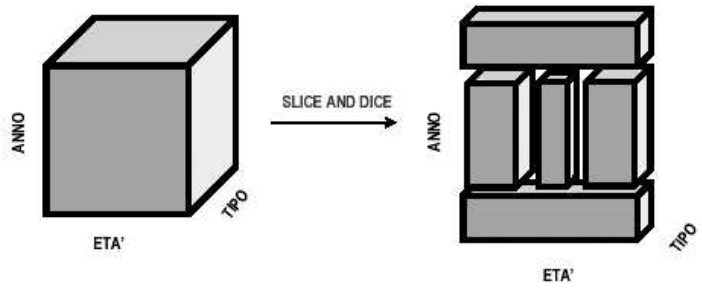
Gerarchia associata alla dimensione "Distribuzione"

Modello Multidimensionale

Operazioni tipiche previste per manipolare i dati in un modello logico multidimensionale sono:

- ***Slice***: è l'operatore che permette di vedere il cubo *trasversalmente* (letteralmente "a fette"), fissando un valore per almeno una delle dimensioni e analizzando i dati relativamente a tutte le altre, cioè concentrando l'attenzione su un ipercubo (n-1) dimensionale del cubo n-dimensionale (*contrazione dimensionale*)
- ***Dice***: è l'operatore per cui fissato un intervallo su ciascuna dimensione, si analizza una *riduzione volumetrica*, senza contrazioni del numero di dimensioni.

Modello Multidimensionale



First page

Modello Multidimensionale

- ***Drill-down:*** è l'operatore che consente di scendere nel dettaglio lungo una o più dimensioni gerarchiche.

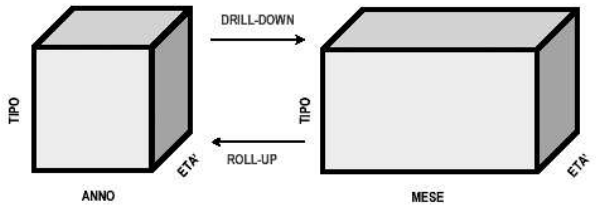
Esempio: mediante un'operazione di drill-down è possibile passare da un'analisi delle vendite per provincia ad un'analisi più particolareggiata, distinguendo in base alle differenti città.

Questo operatore è utile quando si vuole analizzare una causa o un effetto per qualche fenomeno osservato nei dati aggregati.

Modello Multidimensionale

- ***Roll-up*** o ***consolidation*** o ***drill-up***: è l'operatore duale del drill-down, in quanto consente di risalire lungo una o più dimensioni gerarchiche.

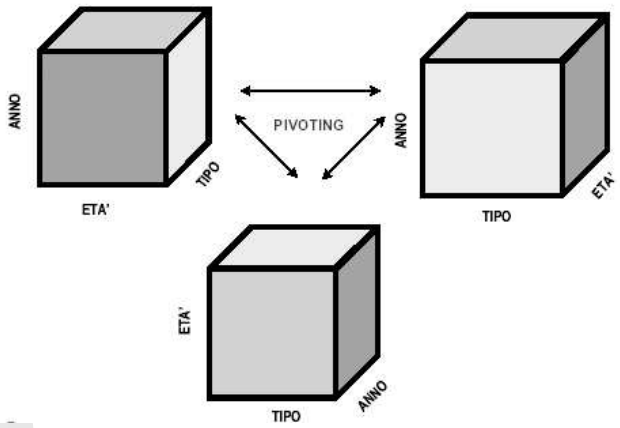
Esempio: partendo dall'analisi di un particolare prodotto si potrebbe passare all'analisi di un'intera gamma di prodotti.



Modello Multidimensionale

- ***Drill-across***: estensione dell'operatore di drill-down, che consente di scendere nel dettaglio contemporaneamente su più dimensioni.
- ***Rolling*** o ***pivoting***: consente di riorientare la vista multidimensionale dei dati, ovvero di poter cambiare la dimensione di analisi. Se lo spazio di analisi è m-dimensionale, sono possibili *m*! prospettive diverse di analisi dei dati.

Modello Multidimensionale



First page

Schema di una DW

In un'applicazione di DW sarebbe naturale tradurre un modello concettuale multidimensionale in un modello *logico* con proprietà analoghe.

Questo tuttavia presuppone l'uso di un DBMS che supporti tale modello logico, cioè un ***MDDBMS*** (***Multidimensional DBMS***), che per l'appunto memorizzano dati numerici o quantitativi categorizzati su diverse dimensioni qualitative.

Esempio: un database multidimensionale (***MDDB***) può memorizzare i dati (quantitativi) relativi alle vendite per diverse linee di prodotto, in diverse città, e per ciascun mese (tre dimensioni qualitative).

Schema di una DW

Gli MDDBMS non sono nuovi. Per circa vent'anni, il pacchetto software EXPRESS della IRI Software Inc. (Burlington, MA), ora di proprietà della Oracle, ha incluso un MDDB. Dagli inizi degli anni '90, molte altre società software hanno prodotto sistemi per MDDB.

- *Vantaggio*: gli MDDB sono ottimizzati per velocizzare e semplificare le interrogazioni, grazie a operazioni di pre-elaborazione.
- *Svantaggio*: Hanno problemi di scalabilità. Un file di 200MB in ingresso a un MDDBMS può occupare fino a 5GB per via delle pre-elaborazioni compiute

Schema di una DW

A questi problemi di natura tecnica, si aggiungono la disponibilità presso le aziende di solo basi di dati relazionali, e la competenza limitata al modello relazionale da parte del personale tecnico che cura il sistema informativo aziendale preesistente.

Tutto ciò porta a prendere in considerazione la possibilità di trasformare un modello concettuale multidimensionale in un modello logico relazionale.

Naturalmente la scelta di un DBMS relazionale per la modellazione di un Data Warehouse, porta ad affrontare i problemi di simulazione di un approccio multidimensionale.

Schema a stella

Lo *schema a stella* (*star schema*) rappresenta il modo più semplice per simulare, mediante l'utilizzo di un database relazionale, un approccio multidimensionale.

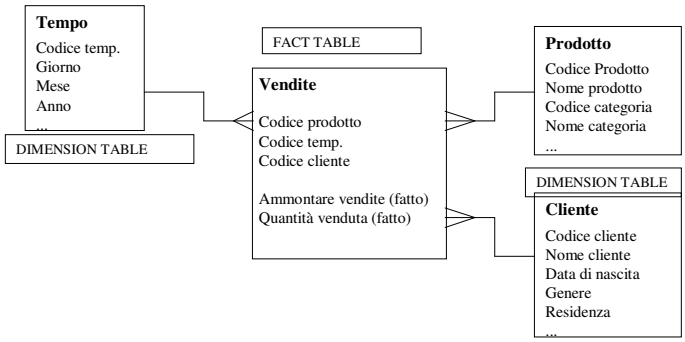
Il nome deriva dall'apparenza del modello dei dati, con una grande tabella centrale circondata da molte tabelle subordinate, in una configurazione a stella.

La tabella centrale, nota come ***tabella dei fatti*** (***fact table***) contiene i dati (numerici) di una cella dell'ipercubo del modello multidimensionale e le chiavi che collegano i dati alle relative dimensioni.

Schema a stella

Le tabelle subordinate, dette anche ***tabelle dimensionali*** (***dimension table***) o ***satellite***, contengono gli attributi che descrivono le componenti dei dati: ce ne sono tante quante sono le dimensioni di analisi individuate nella modellazione concettuale.

Schema a stella



• I fatti sono in forma normale di Boyce-Codd, in quanto ogni attributo non chiave dipende funzionalmente dalla sua unica chiave

• Le dimensioni sono invece relazioni non normalizzate. Ad esempio, in Prodotto, il NomeCategoria dipende dal Codice Categoria, che non è chiave

• Vi sono tre vincoli d'integrità referenziale, uno per dimensione.

Schema a stella

La denormalizzazione di ciascuna delle tabelle dimensionali consente di ridurre il numero di join necessario per risolvere una query, e quindi, in ultima analisi, consente di aumentare l'efficienza dei sistemi di analisi che trasformeranno le interrogazioni formulate dai decision maker per mezzo di interfacce grafiche, in interrogazioni SQL.

Lo scotto da pagare per il guadagno di efficienza è rappresentato dallo spreco di memoria dovuto alla ridondanza dei dati. Ad esempio, il nome della categoria sarà ripetuto per tutti i prodotti che ricadono nella medesima categoria.

Schema a stella

Fortunatamente le anomalie di cancellazione, modifica e aggiornamento, che rappresentano inconvenienti tipici derivanti dalla ridondanza, non si presentano in questo caso semplicemente perché, come accennato precedentemente, i dati presenti nel DW sono per natura statici.

Schema a stella

I dati presenti nella tabella dei fatti, detti ***misure***, devono necessariamente essere numerici. Infatti il fine è sempre quello di estrapolare dati numerici sintetici, tipo “l’incasso totale dovuto alle vendite di un determinato prodotto in una certa area geografica”.

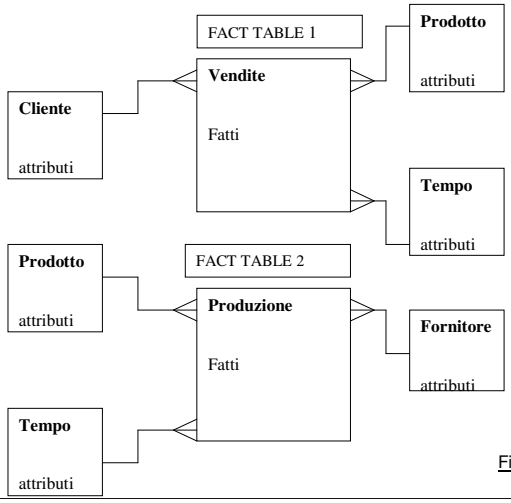
Schema a costellazione

- Lo schema concettuale multidimensionale di un DW può contenere più ipercubi. Questo vuol dire che abbiamo vari gruppi di misure, ognuno dei quali corrisponde a un insieme di dimensioni.
- La traduzione di uno schema concettuale multidimensionale di questo genere in uno schema logico relazionale può portare a più tabelle dei fatti.
- È possibile, inoltre, che le fact table condividano alcune dimensioni. In questi casi è necessario fare una scelta.

First page

Schema a costellazione

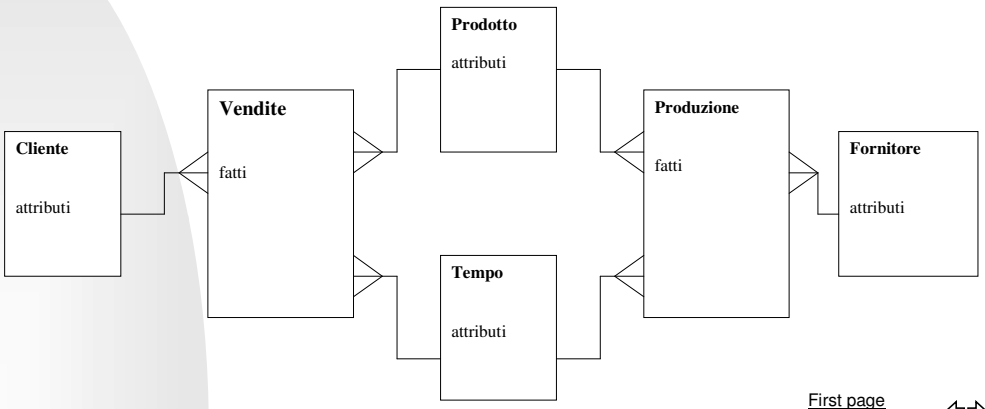
È possibile mantenere separate le tabelle dei fatti con le relative tabelle dimensionali.



First page

Schema a costellazione

Oppure si può costruire uno schema a costellazione nel quale le tabelle dimensionali condivise sono inserite solo una volta e collegate a tutte le fact table da esse accessibili.



First page

Schema a costellazione

La seconda alternativa è ovviamente preferibile, in quanto consente di ridurre notevolmente lo spazio di memoria occupata, anche se il diagramma risulta leggermente più complesso.

In questo modo si perde la limitazione, imposta nello schema a stella, secondo la quale ogni tabella dimensionale ha un solo link.

La condivisione di tabelle dimensionali, tuttavia, richiede una conformità o consistenza dei valori degli attributi delle dimensioni tra i due schemi.

Schema a fiocco di neve

Lo schema a fiocco di neve prende il nome dall'aspetto che assumono generalmente i diagrammi E-R relativi a questo schema. Tale aspetto è determinato dalla normalizzazione delle dimensioni.

Esempio. Si supponga che l'informazione sulla residenza di un cliente corrisponda a tre attributi: la città, la regione e lo stato in cui risiede.

I record della tabella dimensionale cliente presenteranno molte informazioni ridondanti, dovute, in questo caso, alla gerarchia definita fra città, regioni e stati.

Per eliminare tale ridondanza si potrebbe normalizzare la tabella clienti suddividendola in quattro tabelle: una per i clienti, una per riportare le informazioni sulle città di residenza, l'altra per le regioni e infine la quarta per gli stati.

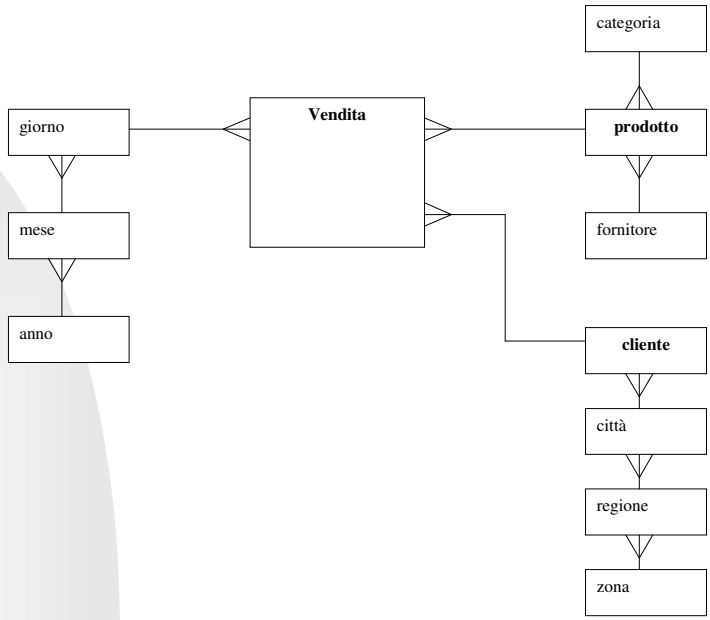
Schema a fiocco di neve

Esempio (cont.)

Diversa è la situazione per la tabella Prodotto, dove oltre a una dipendenza di natura gerarchica con la categoria, c'è una dipendenza di natura non propriamente gerarchica con il fornitore di quel prodotto.

Ancora differente è la situazione nel caso della data, poiché non esiste nessuna dipendenza funzionale fra giorno, mese ed anno. In tal caso, comunque, è possibile definire una sorta di ordine gerarchico fra gli attributi giorno, mese e anno, e si può comunque dividere la tabella Tempo in più tabelle.

Schema a fiocco di neve



First page

Schema a fiocco di neve

In generale, la normalizzazione consente di dividere i dati in funzione delle gerarchie individuate per delle dimensioni.

Ciascuna delle tabelle dimensionali contiene dati relativi ad un solo livello gerarchico e un link per il passaggio al livello gerarchico successivo.

La tabella dimensionale relativa al più basso livello gerarchico è collegata alla tabella dei fatti.

Schema a fiocco di neve

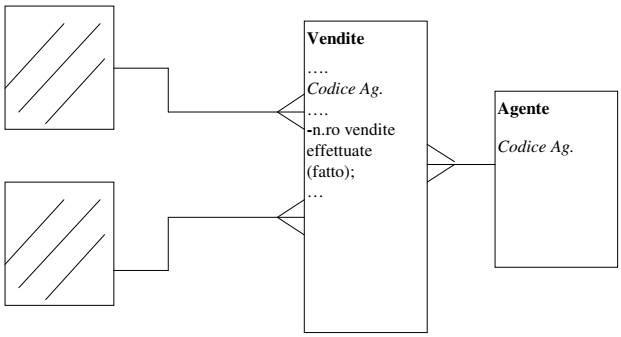
L'utilizzo di uno schema a fiocco di neve è generalmente sconsigliato, poiché il guadagno in termini di occupazione di memoria, dovuto alla normalizzazione, non è solitamente sufficiente a compensare la perdita di efficienza, dovuta al maggior numero di join necessario per risolvere le interrogazioni.

Inoltre nella maggior parte dei casi, il guadagno di occupazione di memoria è poco significativo se si considera la proporzione con la dimensione dell'intero DW, e inoltre si osserva che le dimensioni delle tabelle dimensionali sono irrisorie se confrontate con le dimensioni della tabella dei fatti.

Schema a fiocco di neve

Lo schema a fiocco di neve consente anche di rappresentare le relazioni molti-a-molti, superando così uno dei limiti dello schema a stella.

Esempio: in questo schema a stella una vendita può essere stata seguita da un solo agente e che un agente può seguire differenti vendite.



Schema a fiocco di neve

Esempio (cont.): Tuttavia, in alcuni ambienti, potrebbe accadere che una vendita è seguita da più agenti. Lo schema a stella pertanto non risulta più sufficiente.

È possibile scegliere tra due differenti alternative.

First page

Schema a fiocco di neve

1. l'inserimento nella tabella dei fatti di un record per ognuno degli agenti impegnati nella vendita, il che porta all'inserimento, per ogni vendita, di tanti record identici quanti sono gli agenti impegnati in tale vendita, ad eccezione del codice dell'agente, che ovviamente sarà differente.

- In questo caso, però, si andrebbero a considerare più volte le stesse misure, falsando i risultati di somma, media o di qualsiasi altra operazione di aggregazione.
- Per ovviare a questo inconveniente sarà necessario assicurarsi che, in fase di interrogazione, la query formulata consideri solo una volta ogni vendita per mezzo di controlli sulla chiave primaria della vendita.

Schema a fiocco di neve

2. Utilizzo di tabelle intermedie che, come in un comune modello ER, consentano di scomporre la relazione molti-a-molti in due relazioni uno-a-molti.

First page

Schema a fiocco di neve

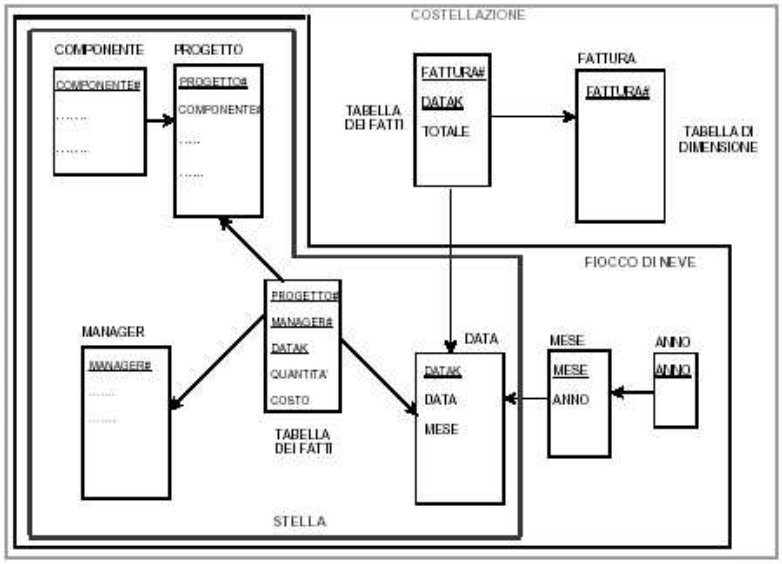
- Lo schema a fiocco di neve è inoltre utilizzato per rappresentare i dati non analitici o di dettaglio.
- Spesso questi non sono presi in considerazione nella progettazione del DW. Tuttavia, questo comporta la necessità di effettuare delle interrogazioni sul sistema operazionale quando invece si rende necessario conoscerli.
- Per ovviare a tale inconveniente, è possibile inserire una *shadow table* collegata alla tabella dimensionale.
- Solo la shadow table conterrà i dati non analitici e questo consentirà di non ridurre le prestazioni del DW.

Schema a fiocco di neve

- Infatti, una qualsiasi interrogazione, che non richiede dati analitici, accederà direttamente alla tabella dimensionale come punto d'ingresso per la tabella dei fatti senza attraversare la shadow table. Si accederà invece a queste ultime solo quando si richiederanno dati non analitici.

Esempio: il numero di telefono di un cliente è considerato un dato non analitico e solitamente non viene preso in esame nel processo di analisi. Tuttavia, l'analista potrebbe voler conoscere l'importo totale delle vendite di un determinato prodotto acquistato dai clienti che hanno come prefisso telefonico lo 06.

Riepilogando …



La dimensione tempo

- Dal punto di vista del modello dimensionale della DW, ci si potrebbe chiedere quale sia la reale utilità di disporre nello schema dei dati di una dimensione temporale esplicita: basterebbe includere la data fra le chiavi della tabella dei fatti, e quindi utilizzare la normale semantica del linguaggio SQL sui data-type per selezionare un mese piuttosto che un anno.
- Tale osservazione è senz'altro valida finché l'uso della variabile temporale resta vincolato ai valori di giorno, mese e anno, nel qual caso la tabella dimensionale tempo sarebbe effettivamente ridondante.

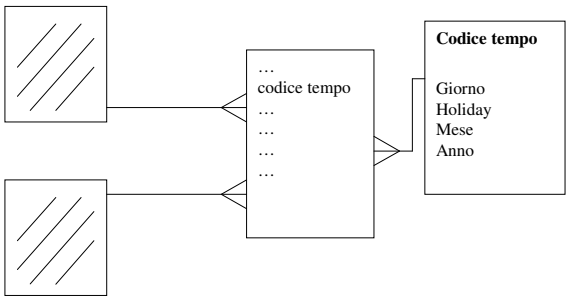
La dimensione tempo

Tuttavia le interrogazioni potrebbero essere più particolareggiate.

Esempio: individuare tutti i dati relativi all'ultimo sabato di ogni mese, o considerare solo i giorni feriali, o ancora considerare solo la prima settimana dei soli mesi estivi.

- In ciascuno di questi casi la memorizzazione della semplice data come uno degli attributi della tabella dei fatti non è ovviamente sufficiente.
- Per tali ragioni, in sistemi DW si utilizza sempre un'apposita tabella dimensionale.

La dimensione tempo



Si noti la separazione tra anni, mesi e giorni, nonché l'utilizzo di opportuni flag per indicare condizioni particolari come il fatto che un giorno sia festivo o meno. In questo modo la selezione, ad esempio, dei soli giorni festivi viene fatta selezionando dalla dimensione *Tempo* solo quei record che presentano il flag *Holiday* opportunamente impostato.

Il progetto Datalight
www.di.uniba.it/~malerba/activities/datalight/

"Datalight: uno strumento di innovazione per le Piccole e Medie Imprese in Puglia" (POP Puglia)

- **Olivetti Ricerca gruppo Getronics**
- **Dipartimento di Informatica Università di Bari**
- **Dipartimento di Progettazione e Produzione Industriale Politecnico di Bari**

Le attività di progetto prevedevano:

- **Analisi dei dati e delle procedure di un campione di Aziende agro-alimentari**
- **Progettazione della soluzione "light" di dw**
- **Progettazione, integrazione e/o implementazione di applicazioni a supporto delle decisioni**
- **Sperimentazione della soluzione**

First page

Progettazione della soluzione “light” di data warehousing

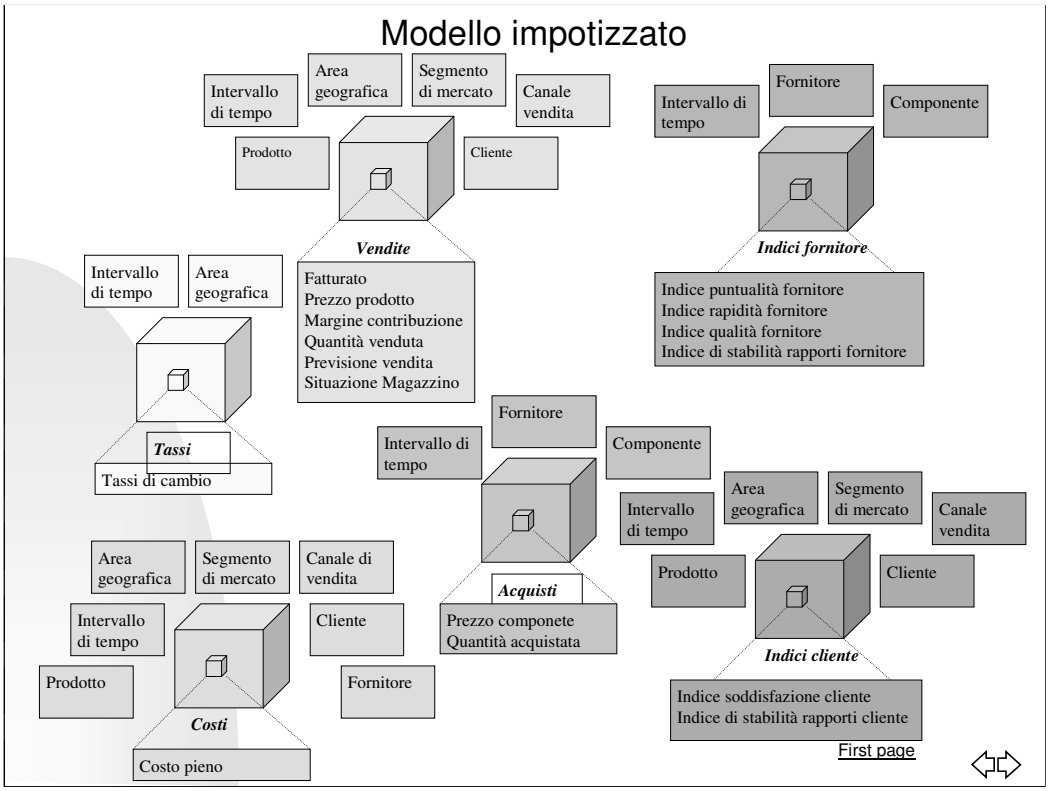
- definizione dell'architettura
- modellazione dei dati → UNIBA
- progettazione di dettaglio

First page

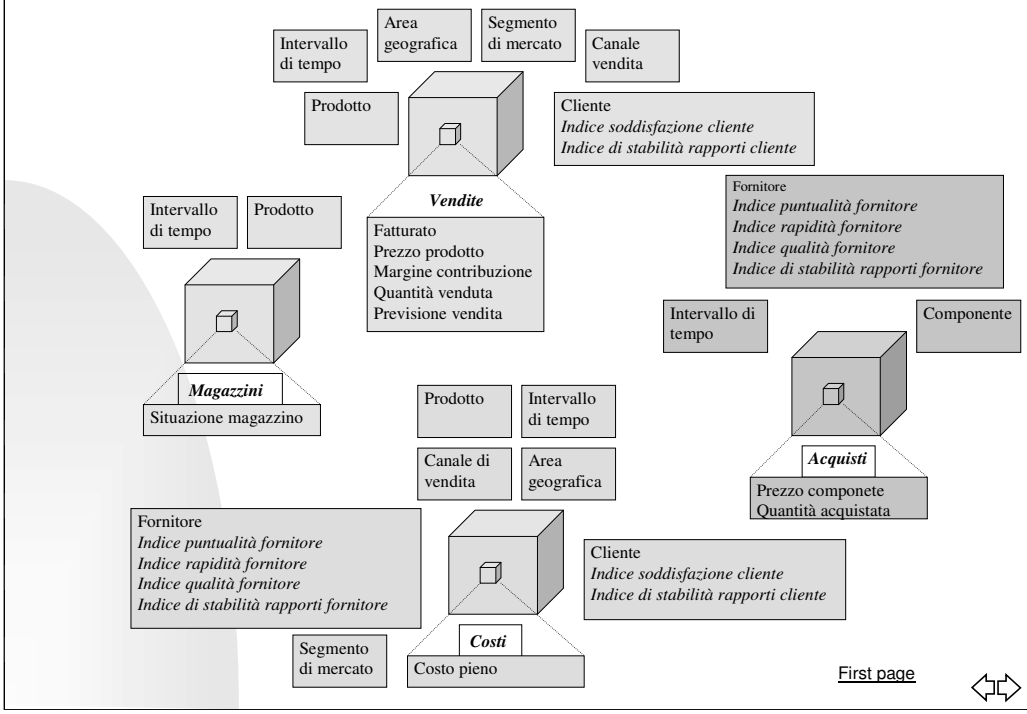
Modellazione dei dati

- Siamo partiti da un modello ipotizzato ricavato da un questionario impostato con i colleghi del Politecnico
- Abbiamo poi provveduto a una semplificazione dopo aver analizzato di dati risultanti dall'indagine.

First page


Modello impotizzato
Intervallo di tempo
Area geografica
Segmento di mercato
Canale vendita
Prodotto
Cliente
Vendite
Fatturato
Prezzo prodotto
Margine contribuzione
Quantità venduta
Previsione vendita
Situazione Magazzino
Intervallo di tempo
Fornitore
Componente
Indici fornitore
Indice puntualità fornitore
Indice rapidità fornitore
Indice qualità fornitore
Indice di stabilità rapporti fornitore
Intervallo di tempo
Area geografica
Tassi
Tassi di cambio
Fornitore
Intervallo di tempo
Componente
Acquisti
Prezzo componete
Quantità acquistata
Area geografica
Segmento di mercato
Canale vendita
Intervallo di tempo
Prodotto
Cliente
Indici cliente
Indice soddisfazione cliente
Indice di stabilità rapporti cliente
Area geografica
Segmento di mercato
Canale di vendita
Intervallo di tempo
Cliente
Prodotto
Fornitore
Costi
Costo pieno
First page

Modello semplificato



On-Line Analytical Processing (OLAP)

- Term introduced by E.F. Codd (1993) in contrast to On-Line Transaction Processing (OLTP)
- The OLAP Council's definition:

"A category of software technology that enables analysts, managers and executives to gain insight into data through fast, consistent, interactive access to a wide variety of possible views of information that have been transformed from raw data to reflect the real dimensionality of the enterprise as understood by the user"

First page

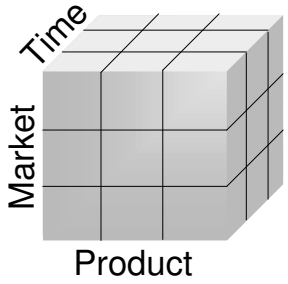
Il termine OLAP fu introdotto da Codd, il papà del modello relazionale, nel 1993, e venne contrapposto al termine OLTP (On-line transaction processing).

Tipiche applicazioni OLTP sono quelle che automatizzano i processi operativi di un'organizzazione.

L' OLAP Council identifica con il termine OLAP una categoria di tecnologia software che consente agli analisti, ai manager e ai dirigenti di intuire ciò che è nei dati attraverso un accesso rapido, consistente e interattivo a una gamma di possibili viste dell'informazione. Tali viste sono state ottenute attraverso trasformazione di dati grezzi e riflettono le reali dimensioni dell'azienda rispetto al punto di vista dell'utente.

On-Line Analytical Processing (OLAP)

- Basic idea: users should be able to manipulate enterprise data models across many dimensions to understand changes that are occurring.
- Data used in OLAP should be in the form of a multi-dimensional cube.



First page

L'idea alla base di questa definizione è che i dati informazionali devono essere modellate mediante un cubo multidimensionale, che consente una visualizzazione molto pragmatica delle connessioni tra i dati.

Un manager descriverà la propria attività con una frase del tipo:

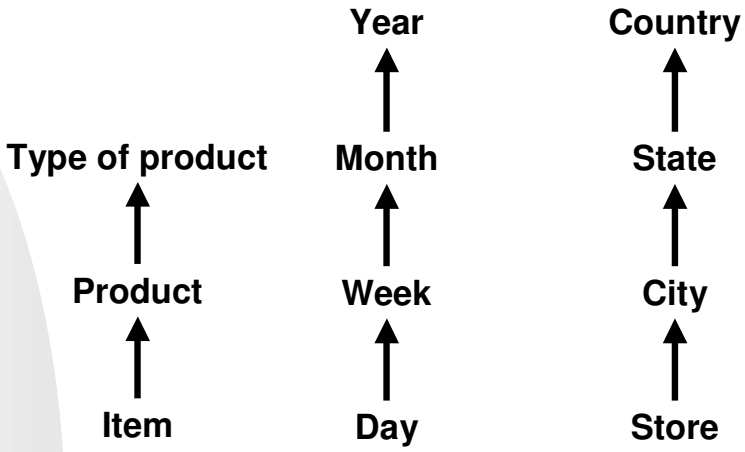
Noi vendiamo prodotti su vari mercati e valutiamo le nostre performance rispetto al tempo

Questa descrizione potrà essere rappresentata come una sorta di cubo, i cui lati etichettati riproducano le variabili di interesse.

Questo modello può apparire fin troppo semplice. Una modellazione entità-relazioni del business rivelerebbe molti dettagli in più circa le relazioni fra le variabili coinvolte, ma ciò non contribuisce alla comprensibilità del business stesso.

Dimensional Hierarchies

- Each dimension can be hierarchically structured

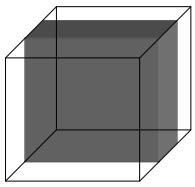


First page

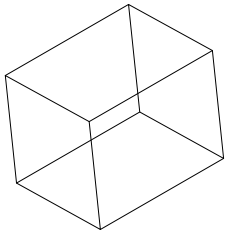
Ciascuna dimensione può essere organizzata in una gerarchia che rappresenta i possibili livelli di aggregazione dei dati.

OLAP Operations

- *Rollup*: decreasing the level of detail
- *Drill-down*: increasing the level of detail
- *Slice-and-dice*: selection and projection



- *Pivot*: re-orienting the multidimensional view of data



First page

Tipiche operazioni OLAP sono:

• rollup: passaggio a livelli di aggregazione superiori

• drill-down: incremento del livello di dettaglio

• slice-and-dice: combinazione di selezione e proiezione

• pivot: rotazione del cubo multidimensionale

Implementing Multi-dimensionality

- *Multi-dimensional databases* (MDDB)
- To make relational databases handle multidimensionality, two kinds of tables are introduced:
  - ***Fact table*: contains numerical facts. It is long and thin.**
  - ***Dimension tables*: contain pointers to the fact table. They show where the information can be found. A separate table is provided for each dimension. Dimension tables are small, short, and wide.**

First page

Per implementare questo modello multidimensionale si può ricorrere o a database specifici, detti appunto multidimensionali, come quello commercializzato da 20 anni dalla EXPRESS software, oppure a database relazionali.

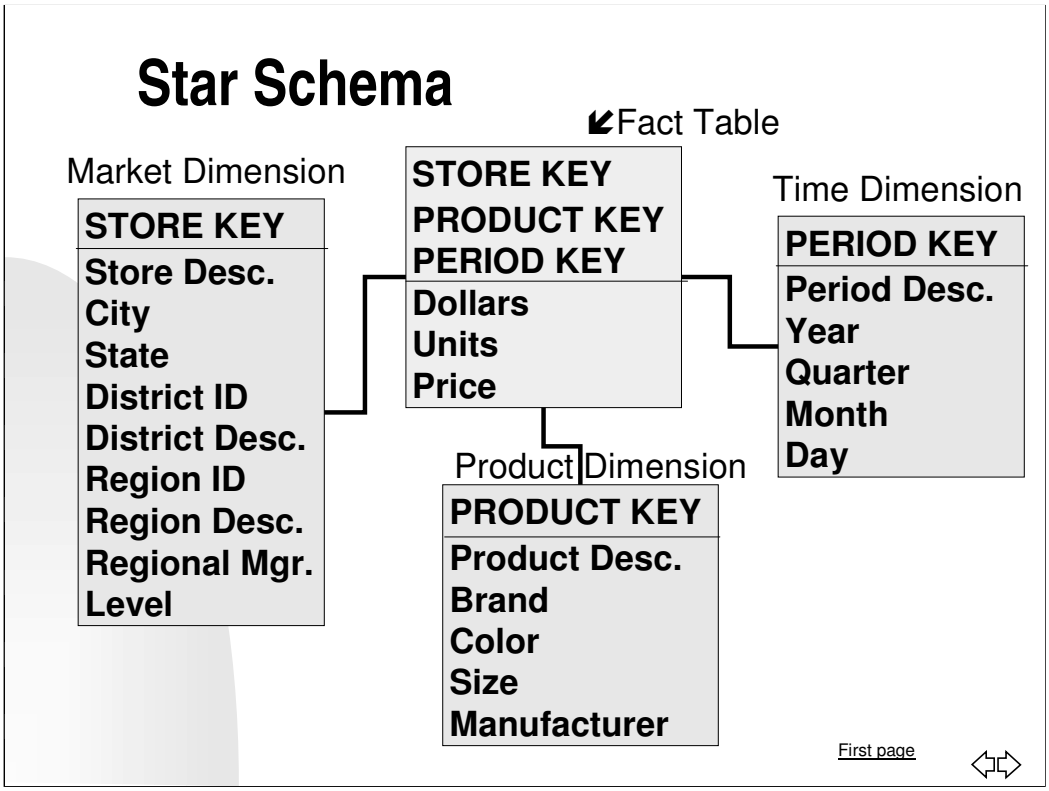
In tal caso il modello multidimensionale verrà rappresentato da due tipi di tabelle:

•una tabella dei fatti, dove vengono memorizzate le misure numeriche del business, ognuna delle quali rappresenta l'intersezione di tutte le dimensioni e quantifica una o più variabili di business.

•Più tabelle dimensionali, che contengono i puntatori alla tabella dei fatti, in modo da riflettere il più naturalmente possibile il modo in cui l'analista vede il business.

La tabella dei fatti è l'unica normalizzata, includendo un'unica combinazione delle chiavi degli elementi dimensionali. Per questo potrebbe contenere qualche centinaio di milioni di record. Basti pensare alla memorizzazione giornaliera di 5000 prodotti per 500 magazzini per un periodo complessivo di 2 anni.

Le tabelle dimensionali, invece, sono generalmente denormalizzate (si osservino i campi codice e descrizione nella prossima slide) e contengono pochi record



Questa implementazione del modello dimensionale su database relazionale prende il nome di *star schema*.

MOLAP, ROLAP, DSS

- The OLAP technology is considered an *extension* of the original DSS technology.
- DSS applications are tools that access and analyze data in relational database (RDB) tables.
- OLAP tools access and analyze multidimensional data (typically three, up to ten-dimensional data).
- OLAP technology is called *MOLAP/ROLAP* (multidimensional/relational OLAP) if it uses an MDDB/RDB.

First page 

Spesso i termini DSS e OLAP sono utilizzati in modo interscambiabile. In effetti sono entrambe tecnologie per l'interrogazione di basi di dati e analisi dei dati, tuttavia la sottile differenza sta proprio nel modello dei dati che *tradizionalmente* essi hanno adottato, ovvero relazionale nel caso dei DSS e multidimensionale nel caso di OLAP.

Inoltre altri due termini ormai in voga sono ROLAP e MOLAP, che stanno semplicemente ad indicare se il modello dimensionale dei dati è stato implementato su base di dati relazionale (RDB) o multidimensionale (MDDB).

OLAP/DSS

- OLAP tools focus on providing multi-dimensional data analysis, that is superior to SQL in computing summaries and breakdowns along many dimensions.
- OLAP tools require strong interaction from the users to identify interesting patterns in data.
- An OLAP tool evaluates a precise query that the user formulates.

First page

Data Warehouse ⇔ Data Mining

⇒ The rational to move from the data warehouse to data mine arises from the need to increase the leverage that an organization can get from its existing warehouse approach.

⇐ After implementing a data mining solution, an organization could decide to integrate the solution in a broader data-driven approach to business decision making. The data warehouse will provide an excellent vehicle for such an integration.

First page

Il passo dal data warehousing al data mining lo stanno facendo le organizzazioni con una vista d'avanguardia sui sistemi di supporto alle decisioni. Il passo è naturale perché:

• una organizzazione che ha gia' investito in un data warehouse conosce il valore strategico dei dati d'impresa ed mostra quindi maggiori aperture sul data mining.

• molto del lavoro piu' pesante, come quello di raccogliere e ripulire i dati è già stato fatto, sicché è pronta a capitalizzare il proprio investimento sul data warehouse.

• c'è una crescente consapevolezza nei business manager e IT manager dell'importanza dei dati raccolti nei vari processi produttivi o operativi. La trasformazione di questi dati in informazioni è un'occasione da cogliere al più presto, per poter supportare adeguatamente i processi decisionali o di governo. Ad esempio, uno studio della Idc (The foundation of wisdom: a study of the financial impact of Data Warehousing), condotto nel 1996 su 62 progetti di Data Warehousing a livello di impresa, mostra un ROI medio pari al 321%.

Tuttavia va riconosciuto che un progetto di data mining che dimostra il valore valore strategico dei dati disponibili in azienda può diventare esso stesso un veicolo per promuovere iniziative dal budget chiaramente più pesante, come quella di sviluppare un data warehouse. Un data warehouse integrato con una soluzione di data mining rappresenta la componente chiave di una moderna infrastruttura di decision-making.

Critical Success Factors for Business Applications

- People
  - **Find a sponsor for the application**
  - **Select the right user group**
  - **Involve a business analyst with domain knowledge**
  - **Collaborate with experienced data analysts**
- Data
  - **Select relatively clean sources of data**
  - **Select a limited set of data sources (e.g., the data warehouse)**

First page

Critical Success Factors for Business Applications (cont.)

- Application
  - **Understand business objectives.**
  - **Analyze cost-benefits and significance of the impact on business problem.**
  - **Consider legal or social issues in collecting input data**

First page