9- Basi di dati direzionali

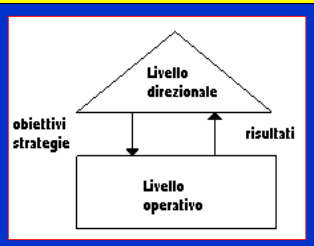
Basi di Dati
per la gestione dell'Informazione

A. Chianese, V. Moscato, A. Picariello,
L. Sansone

SOMMARIO

- Sistemi Informativi Direzionali (SID)
- Architettura dei SID
- La base di dati direzionale (Data Warehouse, DWH)
  - caratteristiche
  - architettura
  - processi di acquisizione e gestione dei dati
- Tecnologie di Data Warehousing
  - query & reporting
  - analisi multidimensionale
  - Knowledge Discovery in Database e Data mining
- Modelli relazionali di DWH
- Progetto di un DWH:
- Esempio: statistiche vendite farmaci

SISTEMI INFORMATIVI DIREZIONALI



Il ***livello direzionale*** si occupa di quelle attività necessarie alla definizione degli **obiettivi** da raggiungere ed alle **strategie** da intraprendere per perseguirli.

Il ***livello operativo*** viceversa si occupa delle attività attraverso cui l'azienda produce i propri servizi e/o prodotti fornendo a sua volta al livello direzionale informazioni sui risultati raggiunti.

Tipi di sistemi direzionali

pianificazione strategica:
determina gli obiettivi generali dell'azienda

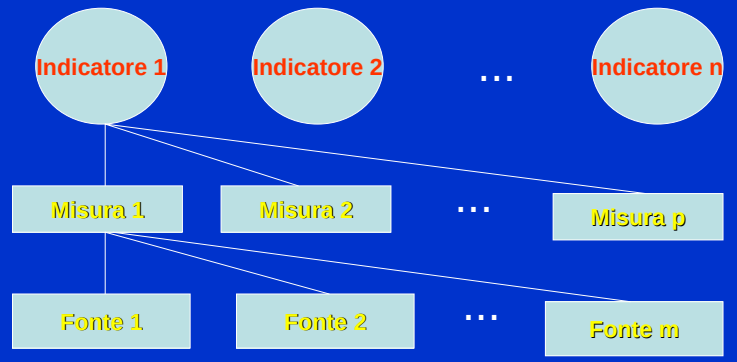
controllo direzionale (di gestione) :
definisce traguardi economici ovvero risultati da conseguire a medio termine e loro verifica

controllo operativo:
assicura che le attività dei processi aziendali procedano nel modo prefissato.

Tipologia delle informazioni direzionali

- I sistemi direzionali "analizzano" informazioni fortemente **aggregate**.
- Essi devono infatti fornire ai dirigenti aziendali "**informazioni sintetiche**" dette **indicatori prestazionali** – ricavati a partire da misure gestionali quali quantità vendute, ricavi globali - che possano quantizzare il perseguimento dei loro obiettivi ed in genere l'andamento dell'azienda.

Il paradigma "indicatori – misure – fonti"

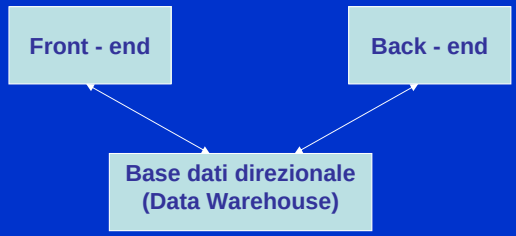


Dimensioni (variabili) di analisi

L'analisi delle **"informazioni direzionali"** deve essere condotta in diverse **dimensioni**:

- la dimensione ***tempo*** ;
- la *dimensione* ***prodotto*** : finalizzata all'analisi di costi e ricavi (i quali sono evidentemente indicatori contabili o monetari);
- la *dimensione* ***processi*** : finalizzata al controllo di indici di efficienza ed efficacia come la tempestività ;
- la *dimensione* ***responsabilità*** : ad ogni "centro" così come indicato dall'organigramma aziendale possono essere associati alcuni indici che forniscono dei rendiconti sulle prestazioni dei singoli dirigenti;
- la dimensione ***cliente*** : al fine di analizzare redditività, volume di affari e bacino di utenza.

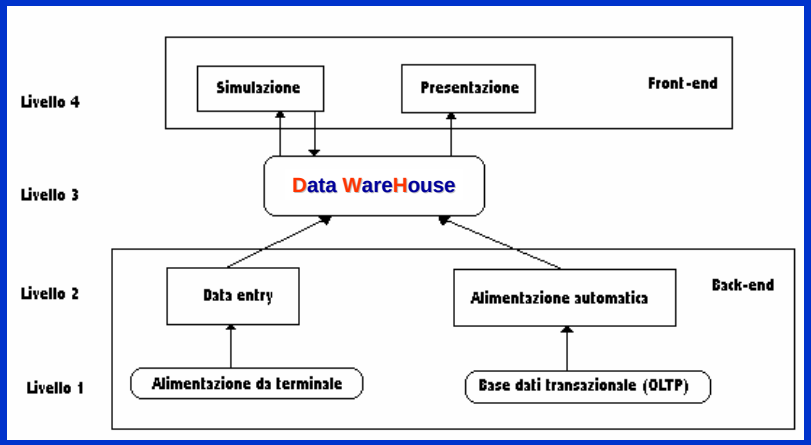
ARCHITETTURA DEI SID



Il sottosistema **front-end** comprende tutte quelle elaborazioni necessarie alla presentazione delle informazioni della base dati direzionale – detta "Data Warehouse" - utili all'utente finale.

Il sottosistema **back-end** provvede ad alimentare automaticamente la Base dati della direzione detta "Data Warehouse" in maniera periodica estraendo le informazioni di interesse per il Sistema Informativo Aziendale.

Dettagli dell'architettura dei SID



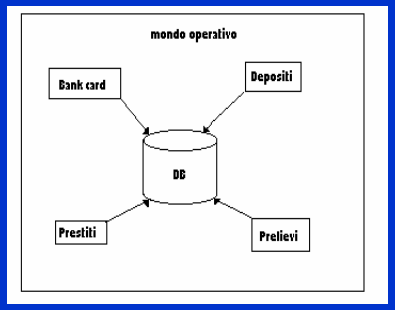
DATA WAREHOUSE: DWH

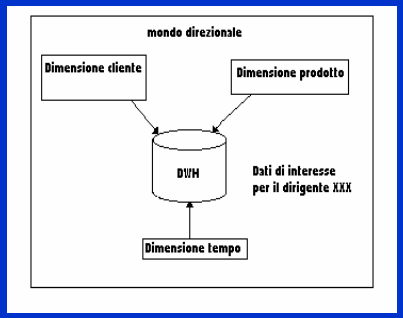
- Una base di dati relazionale utile ai processi direzionali caratterizzata dall'essere:
  - **orientata ai soggetti dell'elaborazione**
  - **integrata**
  - **tempo - variante**
  - **non volatile**

  ricavata dagli ambienti operativi (OLTP) e **fisicamente separata** da essi.

Orientata al soggetto

- E' una **collezione di dati orientata ai soggetti dell'elaborazione:** ciò è in contrasto con le più classiche applicazioni di data base che hanno un orientamento processivo / funzionale:

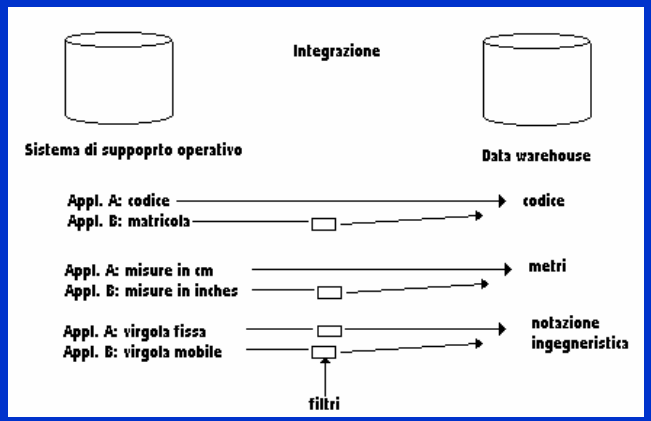




Integrata

- Le informazioni dei **processi operativi** che alimentano un DWH **possono provenire "in pratica" da differenti basi dati** e quindi avere caratteristiche differenti e formati inconsistenti
- **I dati di un DWH devono quindi essere integrati.** Questa integrazione può essere evidenziata in modi differenti quali:
  - misure consistenti delle variabili
  - attributi fisici dei dati
  - strutture di codifica
  - convenzioni sui nomi.

... problematiche di integrazione

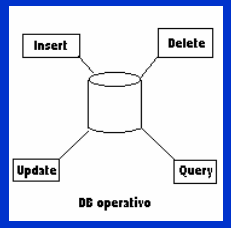


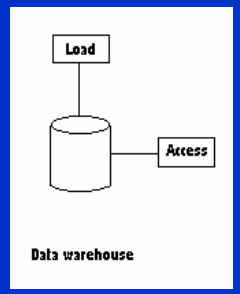
Tempo - variante

- Tutti i dati contenuti in un data warehouse si riferiscono ad un **preciso arco temporale**.
- Un DWH rappresenta dati in genere su un lungo periodo (ad esempio cinque, dieci anni) rispetto ai dati contenuti in un data base operativo validi e consistenti per un periodo molto più corto (ad es. giorni o settimane).
- Ogni struttura di base in un DWH contiene implicitamente o esplicitamente un riferimento ad un valore temporale contenuto nella tabella dei tempi la quale ricordiamo rappresenta per gli utenti direzionali una dimensione di analisi.

Non volatile

- Dal punto di vista delle operazioni consentite su un DWH a differenza dei dati di un data base tradizionale che possono essere inseriti, modificati ed acceduti, i dati di un DWH possono essere "**solo caricati ed acceduti**" poiché essi rappresentano successive istantanee della realtà elaborativa.

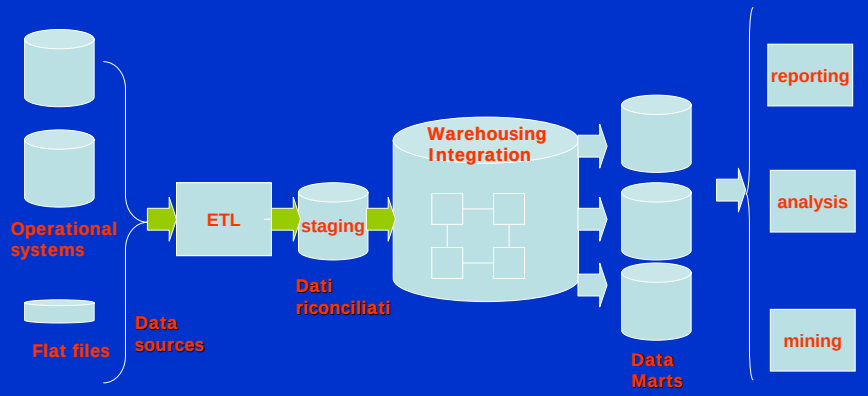




Sistemi "operativi" (OLTP) e sistemi "direzionali"(OLAP)

SISTEMI OPERATIVI	SISTEMI DIREZIONALI
Memorizzano dati correnti	Memorizzano dati storici
Memorizzano dati dettagliati	I dati sono debolmente o fortemente aggregati
I dati sono dinamici	I dati sono pressoché statici
Le elaborazioni sono ripetitive	Effettuano elaborazioni ad hoc
Massimizzano il throughput rispetto alle transazioni	Hanno un throughput di transazioni medio-basso
Sono transaction driven	Sono Analysis driven
Sono orientati alle applicazioni	Sono orientati al soggetto
Supportano le decisioni day-to-day	Supportano decisioni strategiche
Servono un grosso numero di utenti operativi	Servono un numero di utenti manageriali relativamente basso

ARCHITETTURA DI UN DWH



Data sources

- Il DWH utilizza sorgenti di dati eterogenee provenienti da:
  - sistemi operazionali - dati prelevati in ambienti di produzione - di tipo relazionale,
  - applicazioni "legacy" - applicazioni aziendali esistenti che non rispondono a requisiti architetturali moderni e attuali.
  - sistemi informativi esterni strutturati
  - flat files cioè dati non strutturati in tabelle relazionali.

ETL

I dati delle varie sorgenti devono essere opportunamente trattati e cioè:

- ripuliti per eliminare eventuali incongruenze ed inconsistenze,
- completati di parti mancanti,
- integrati secondo uno schema comune.

- A tal fine sono definiti sofisticati strumenti di *Extraction, Transformation and Loading* (**ETL**) che hanno lo scopo di integrare i vari dati provenienti dalle diverse sorgenti informative.

Staging

- I dati integrati, corretti, filtrati e validati sono materializzati in una opportuna area detta di staging che contiene i cosiddetti **dati riconciliati** che:
  - costituiscono un modello di dati comune e di riferimento per l'azienda
  - sono pronti a fornire alimentazione al DWH.

Warehousing Integration

- Le informazioni vengono raccolte in un singolo contenitore (magazzino) “logicamente” centralizzato, che è in senso più stretto il DWH.
- Esso contiene anche i metadati che rappresentano “dati sui dati” della DWH e sono indispensabili per la gestione della DWH stessa.

DWH Integration: Metadati

- I meta-dati rappresentano dati sui dati realmente contenuti in un data warehouse e consentono di individuare:
  - i dati all’interno di un DWH.
  - le informazioni per le operazioni di loading
    - Sorgente dati, cambi da effettuare sui dati prima dell’inserimento nel DWH
  - le informazioni legate al query manager:
    - Informazioni sui profili delle query utilizzate dai gruppi di utenti al fine di migliorare le prestazioni.
  - le informazioni di gestione dei dati:
    - Organizzazione tabelle, viste, indici.

Data Marts: informazioni

- Memorizzano un sottoinsieme dei dati del DWH normalmente in forma molto aggregata.
- Da un punto di vista logico le informazioni sono direzionali e relative ad una **particolare area di business** o un **particolare dipartimento aziendale** (vendita, produzione, …).
- Possono essere autonomi o collegati al DWH centrale. Essendo più piccoli del DWH sono più facilmente gestibili e più efficienti.
- Un data-mart a differenza di un DWH :
  - si focalizza solo sui requisiti di un particolare dipartimento;
  - non contiene dati operativi;
  - contiene meno informazioni e quindi è di più facile navigazione.

Data Marts: struttura

- Gli **approcci per la costruzione dei data mart** possono essere diversi:
  - creare un *data mart come vista a partire da un DWH.*
  - costruire delle *infrastrutture dedicate* ed eventualmente e successivamente integrate col DWH centralizzato.
- Le ragioni che spingono alla costruzione di data mart dedicati sono diverse:
  - consentono agli utenti un accesso rapido alle informazioni più frequentemente usate migliorando i tempi di risposta del sistema (essendo diminuito il volume dei dati da visitare).
  - forniscono strutture dati appropriate ad esigenze specifiche agevolando le tecniche di data mining.
  - Il costo di creazione e gestione di una infrastruttura dedicata è molto più basso rispetto a quello di vista su un DWH

ACQUISIZIONE E GESTIONE DATI

Inflow

- Col termine *inflow* si indicano i processi associati alla **estrazione, filtraggio e caricamento dei dati dai source system OLTP al DWH.**
- I dati prelevati vanno opportunamente riorganizzati secondo gli scopi per i quali il DWH è progettato. Questa riorganizzazione riguarda:
  - la rimozione o l'aggiunta di alcuni campi e la loro normalizzazione;
  - la verifica della consistenza dei dati con i dati già presenti nel DWH.
- I dati da estrarre vengono preventivamente tenuti in una memoria temporanea nella quale sono soggetti ai test di consistenza e alle operazioni di filtraggio opportune.

Upflow

- Col termine *upflow* si indicano i processi associati all' **aggiunta di dati al DWH** attraverso operazioni di aggregazione.
- Queste operazioni tendono a porre i risultati operazionali in **"varie forme"** – dettagliati, parzialmente aggregati, fortemente aggregati – per agevolare gli utenti finali nella costruzione di diagrammi, animazioni, grafi etc.

Downflow

- I processi di *downflow* sono associati ad operazioni relative all' **archiviazione sicura dei dati** nel DWH.
- Questi processi assicurano la ricostruzione dello stato del DWH in caso di guasti o perdita di dati.

Outflow

- Indichiamo con *outflow* l'insieme dei processi demandati a **rendere disponibili i dati all'utente finale**.
- Essi quindi soddisfano le richieste di utente le quali devono essere servite in maniera efficiente.

Metaflow

- I processi di *metaflow* sono associati alla **gestione dei metadati**.

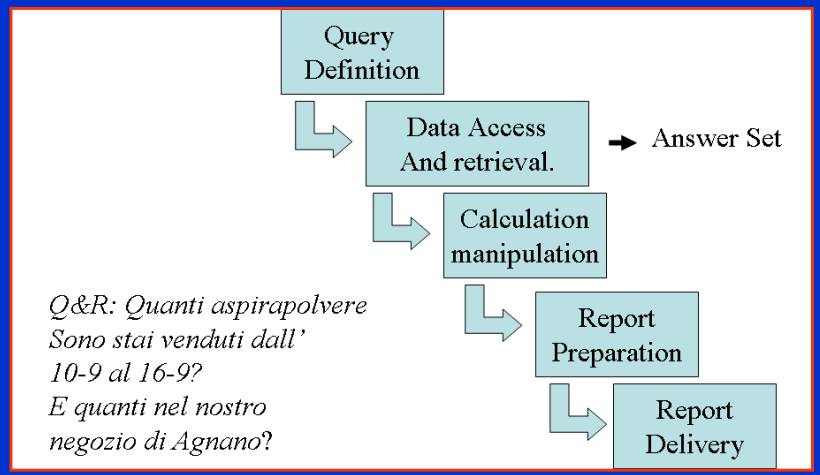
TECNOLOGIE DI "DATA WAREHOUSING"

- I dati integrati nella warehouse devono essere facilmente consultabili ai fini di :
  - stendere report (**query & reporting**)
  - di effettuare analisi e simulazioni avanzate (on line analytical processing,**olap**).
  - individuare regole nascoste nei dati (**K**nowledge **D**iscovery in **D**atabase, **KDD**)

Query & Reporting

- E' il processo di
  - Porre una interrogazione.
  - Rilevare dati fondamentali dal DWH.
  - Trasformare i dati in un contesto appropriato.
  - Preparare i dati in formato leggibile.
  - Spedire i dati in formato leggibile.

Query e reporting



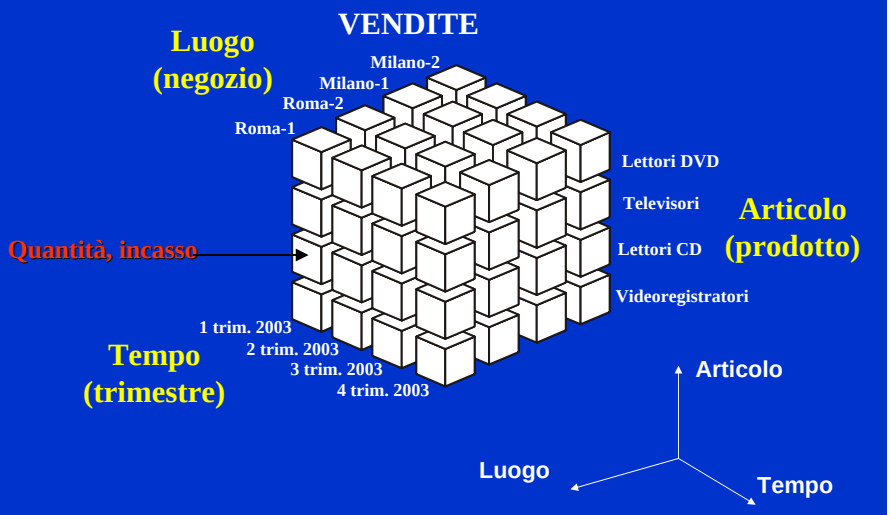
Analisi multidimensionale

- L'analisi multidimensionale dei dati consiste di operazioni interattive di aggregazione e disaggregazione lungo opportune coordinate dette dimensioni - spazio, tempo, prodotto,cliente - di "misure gestionali" relative ad un fissato "indicatore prestazionale" di interesse per la direzione aziendale.

Analisi multidimensionale

- Concetti rilevanti:
  - **fatto** — un indicatore prestazionale sul quale centrare l'analisi (ad. es. vendite)
  - **misura** — una proprietà atomica di un fatto (ad. es. quantità vendute, incasso totale)
  - **dimensione** — rappresenta una "variabile" di interesse dell'analisi (ad. es: articolo, luogo, tempo)
- Ciascuna dimensione è organizzata in una gerarchia che rappresenta i possibili livelli di aggregazione/disaggregazione per le misure gestionali
  - per la dimensione tempo: giorno, mese,trimestre, anno
  - per la dimensione luogo: negozio, città, provincia, regione
  - per la dimensione articolo: descrizione, categoria

Rappresentazione logica multidimensionale



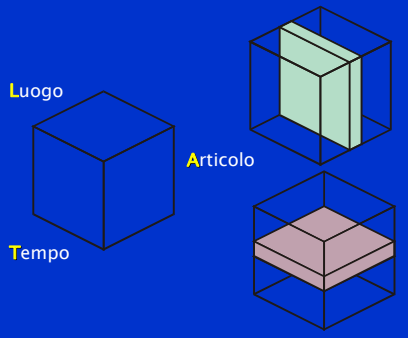
Operazioni su dati multidimensionali

- **Slice & dice** — seleziona e proietta
- **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
- **Pivot** — re-orienta il cubo

Slice and dice

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

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



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

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

Risultato di slice and dice

LETTORI DVD	1 trim. 03	2 trim. 03	3 trim. 03	4 trim. 03
Roma-1	38	91	66	198
Roma-2	155	219	248	265
Milano-1	121	273	266	326
Milano-2	222	122	155	200

... e poi di roll-up

LETTORI DVD	1 trim. 03	2 trim. 03	3 trim. 03	4 trim. 03
Roma	193	310	314	463
Milano	343	395	421	526

Risultato di roll-up

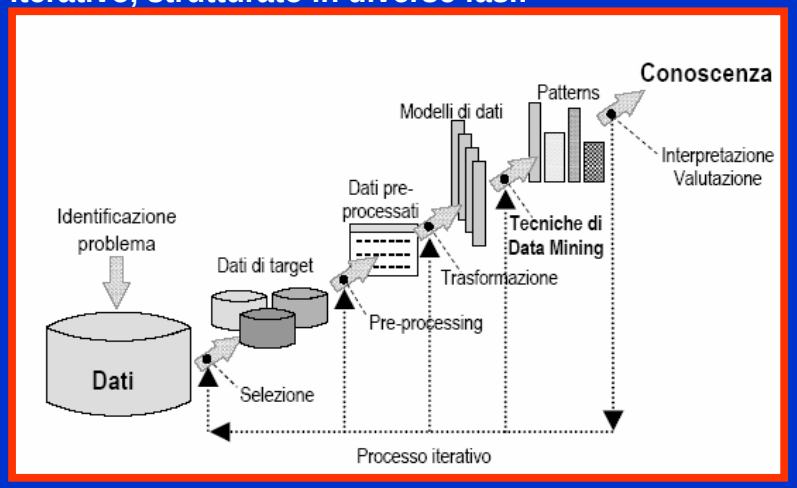
Per tutti i negozi	1 trim. 03	2 trim. 03	3 trim. 03	4 trim. 03
Lettori DVD	536	705	735	989
Televisori	567	716	606	717
Lettori CD	187	155	186	226
Videoregistratori	175	191	202	319

... e poi drill - down

Per tutti i negozi	Gen 03	Feb 03	Mar 03	Apr 03	Mag 03	Giu 03	…
Lettori DVD	165	178	193	205	244	256	…
Televisori	154	201	212	245	255	216	…
Lettori CD	54	88	45	24	65	66	…
Videoregistratori	56	64	55	52	64	75	…

Knowledge Discovery In Database

Il processo di *KDD* è un processo interattivo e iterativo, strutturato in diverse fasi:



Knowledge Discovery in Database (2)

- *Fase 1*: si identifica il problema tenendo conto della relativa conoscenza già acquisita in precedenza e gli obiettivi che si vogliono perseguire.
- *Fase 2*: si seleziona l'insieme dei dati oggetto del processo di estrazione (*discovery*) della conoscenza.
- *Fase 3*: si "puliscono" e si normalizzano i dati attraverso, ad esempio, l'eliminazione dei dati rumorosi e dei valori estremi, la gestione dei campi vuoti …
- *Fase 4*: si individuano le caratteristiche salienti per rappresentare il fenomeno che si sta analizzando in funzione dell'obiettivo definito, tendendo a ridurre il numero delle variabili prese in considerazione.

Knowledge Discovery in Database (3)

- *Fase 5*: si effettua il *data mining*, **cioè si compie la ricerca dei *pattern* d'interesse**.
- *Fase 6*: si interpretano i *pattern* "scoperti" con la possibilità di ritornare alle fasi precedenti per ulteriori iterazioni.
- *Fase 7*: si consolida e si formalizza la conoscenza acquisita mediante realizzazione/integrazione di un sistema applicativo, redazione di documentazione, presentazione alle parti interessate ….

KDD e Data Mining

- In questa ottica, il KDD è un processo non banale di **identificazione dai "dati" di "pattern" validi**, precedentemente sconosciuti, potenzialmente utili ed ultimamente comprensibili.
  - per "dato" si intende in questo contesto un insieme di fatti associati ad una transazione;
  - per "pattern" si intende un **modello** o **insieme di regole** applicabile a un sottoinsieme di dati.
- Cioè per processo di estrazione di un pattern si intende il processo di individuare un modello - sotto forma di un insieme di regole - che ben si adatta ai dati in esame.

... quindi il data mining

- è un approccio **alternativo all'analisi multidimensionale** per estrarre informazioni di supporto alle decisioni da un data warehouse
- talora le tecniche di ricerca di "informazione nascosta" in una collezione di dati **si applica a dati** "**destrutturati**" ad es. **collezioni di transazioni**.

Esempio di data mining

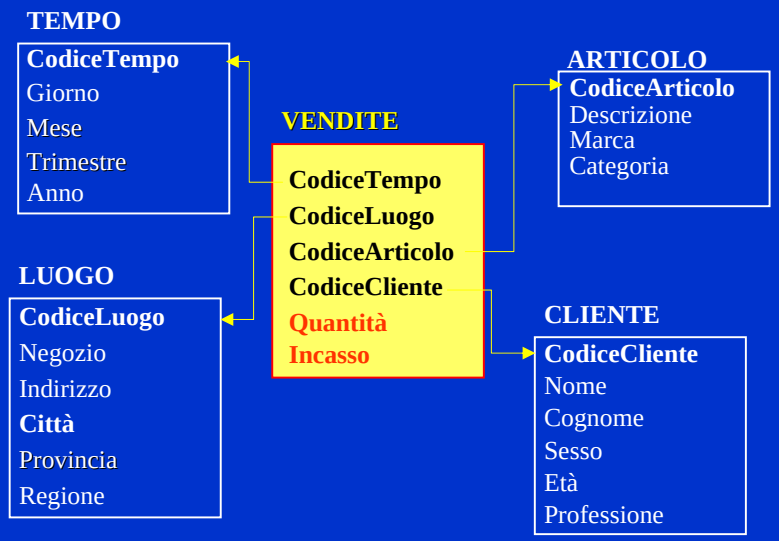
- Trovare delle "**regole**" che consentano di "scoprire" in una transazione (dato) dalla presenza di un oggetto la presenza di un altro oggetto.
- Si parla di regola associativa del tipo premessa $\Rightarrow$ conseguenza, ad esempio:

  Pannolini $\Rightarrow$ Birra
- la verifica della "**scoperta**" di una regola significativa avviene attraverso la valutazione del "supporto" S e della "confidenza" C della regola:
  - S% il (ad. esempio il 2%) tra tutte le transazioni contiene premessa e conseguenza cioè entrambi gli oggetti - **supporto della regola** cioè sua **rilevanza statistica.**
  - C% (ad esempio il 30%) delle transazioni che contiene la premessa (Pannolini) contiene anche la conseguenza (Birra) - **confidenza della regola** cioè sua **forza.**

MODELLI RELAZIONALI DI DWH
Schemi relazionali a stella

- Uno **schema a stella** è composto da
  - una tabella principale, chiamata **tabella fatti**
    - memorizza i fatti e le sue misure
      - Le misure più comuni sono numeriche, continue e additive
  - due o più tabelle ausiliarie, chiamate **tabelle dimensione**
    - una tabella dimensione rappresenta una dimensione rispetto alla quale è interessante analizzare i fatti
      - memorizza i membri delle dimensioni ai vari livelli
      - gli attributi sono solitamente testuali, discreti e descrittivi

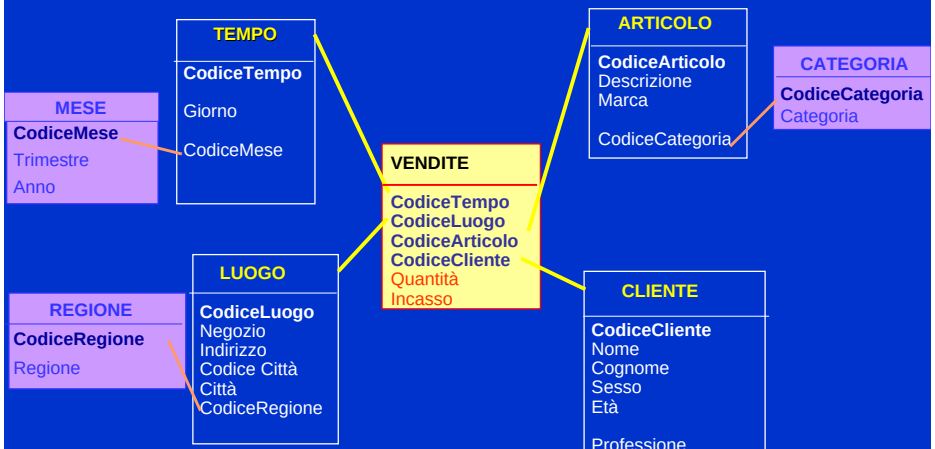
Esempio di schema a stella



Caratteristiche di uno schema a stella

- Nella tabella fatti:
  - la chiave primaria è composta da riferimenti alle chiavi di tabelle dimensione
  - gli altri campi rappresentano le misure.
  - è in 3NF
- Nella tabella dimensioni:
  - la chiave primaria è semplice
  - gli altri campi memorizzano **i livelli** della dimensione
  - tipicamente denormalizzata; se normalizzata o parzialmente normalizzata si ha uno schema a fiocco di neve.

Esempio di schema a fiocco di neve



Additività dei fatti

- Un fatto è normalmente **additivo** cioè ha senso sommarlo rispetto a ogni possibile combinazione delle dimensioni da cui dipende
  - l'incasso è additivo perché ha senso calcolare la somma degli incassi per un certo intervallo di tempo, insieme di prodotti e insieme di negozi
  - l'additività è una proprietà importante, perché le applicazioni del data warehouse devono solitamente combinare i fatti descritti da molti record di una tabella fatti

Formato delle interrogazioni di aggregazione sullo schema a stella

- Le interrogazioni assumono solitamente il seguente formato standard :

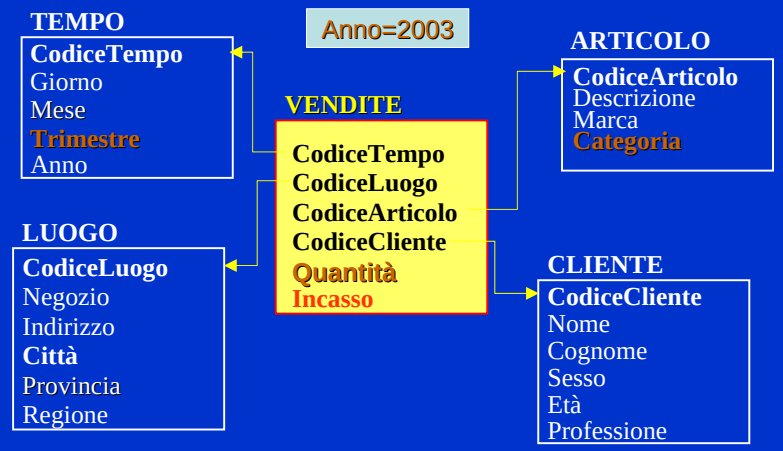
```
SELECT D1.L1,.., Dn.Ln, Aggr1(F.M1),.., Aggrk(F.Ml)
 FROM Fatti as F, Dimensione1 as D1, ..,
      DimensioneN as Dn
 WHERE Join-predicate(F,D1) and ..
     and Join-predicate(F,Dn)
     and selection-predicate
 GROUP BY D1.L1, ..., Dn.Ln
 ORDER BY D1.L1, ..., Dn.Ln
```


Esempio

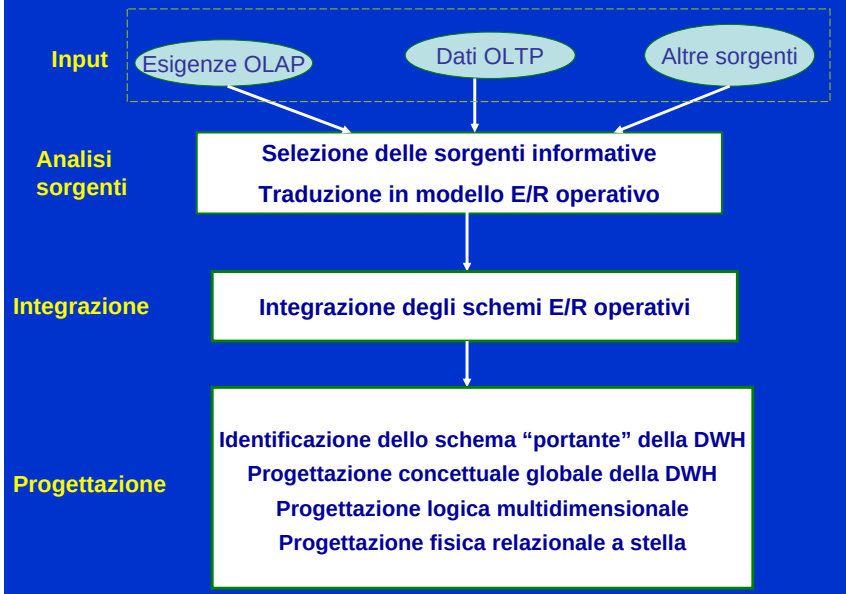
Per ogni categoria di articolo e trimestre per l'anno 2003 calcolare le quantità totali vendute

```
SELECT A.Categoria, T.trimestre, sum(V.Quantita)
FROM Vendite as V, Articolo as A, Tempo as T
WHERE V.CodiceArticolo = A.CodiceArticolo and
      V.CodiceTempo = T.CodiceTempo and T.Anno = 2003
GROUP BY A.Categoria, T.trimestre
ORDER BY A.Categoria, T.trimestre
```


Schema dell'esempio



PROGETTO DI UN DWH



ESEMPIO DI PROGETTO DI UNA DWH:
Statistiche vendite farmaci

Il Ministero della Salute ha commissionato la progettazione di un Data Warehouse per effettuare analisi e statistiche circa le vendite di farmaci da parte delle varie farmacie italiane.

In particolare si vogliono analizzare le statistiche relative alle tipologie di farmaci venduti suddivisi per area geografica e orizzonte temporale, nonché semplici statistiche sull'utenza consumatrice.

Individuazione ed analisi sorgenti informative

- La prima fase nella progettazione del DWH consiste nell'individuazione e analisi delle sorgenti informative contenenti i dati operazionali da analizzare.
- Da un colloquio con il committente, si evince che ogni farmacia utilizza una base di dati operazionale per la gestione delle vendite dei farmaci implementata attraverso un apposito DBMS relazionale.

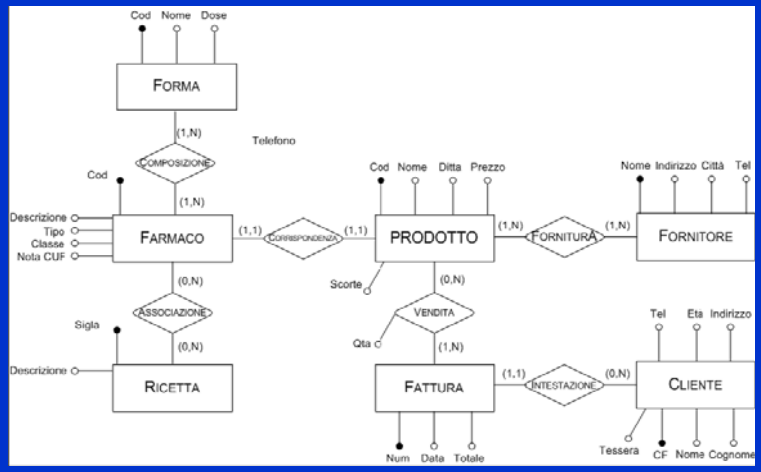
Lo schema della base dati operativa

PRODOTTI (Cod, Nome, Ditta, Prezzo, Scorte)
FORNITORI (Nome, Indirizzo, Città, Tel)
FORNITURE (NomeFornitore:FORNITORI, CodProdotto:PRODOTTI)
CLIENTI (CF, Tessera, Nome, Cognome, Età, Indirizzo, Tel)
FATTURE (Numero, Data, Totale, CFCliente:CLIENTI)
VENDITE (NumFattura:FATTURE, CodProdotto:PRODOTTI, Qta)
FARMACI (Cod, Nome, Tipo, Descrizione, Nota CUF, Classe,
CodProdottto:PRODOTTI)
FORME(Cod, Nome, Dose)
COMPOSIZIONI(CodFarmaco:FARMACI, CodForma:FORME)
RICETTE(Sigla, Descrizione)
ASSOCIAZIONI(CodFarmaco:FARMACI, SiglaRicetta:RICETTE)

Dallo schema logico allo schema E/R

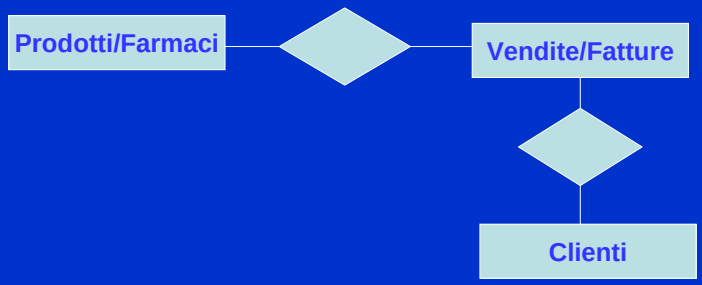
- Dall'analisi di tale schema logico, seguendo poi un semplice processo di *reverse engineering*, è possibile individuare lo schema E/R di riferimento (uguale per le varie sorgenti informative).

Schema E/R della base dati operativa



Derivazione di uno schema concettuale portante della DWH

- Per la progettazione del DWH si identificano in prima istanza le seguenti *informazioni direzionali :*
  - Vendite/Fatture,
  - Prodotti/Farmaci,
  - Clienti.



… con fatti misure e dimensioni

- Lo schema concettuale indica che i prodotti di una farmacia (un tipo di farmaco) sono venduti con apposita fattura ai clienti, di cui la farmacia stessa possiede i dati anagrafici.
- Da tale schema emergono nel contempo:
  - il fatto principale, ovvero la “vendita dei farmaci”;
  - le misure, ovvero il “prezzo dei farmaci” e la “quantità venduta”;
  - alcune delle dimensioni dell’analisi, ovvero i “prodotti” e i “clienti”.

Integrazione degli schemi concettuali

- Nell’ipotesi che le farmacie utilizzano tutte lo **stesso schema logico** dei dati, sarà semplice effettuare le operazioni di integrazione dei dati nella base di dati direzionale, e, quindi il modello concettuale definito precedentemente, può essere considerato come il punto di partenza per l’implementazione del DWH.

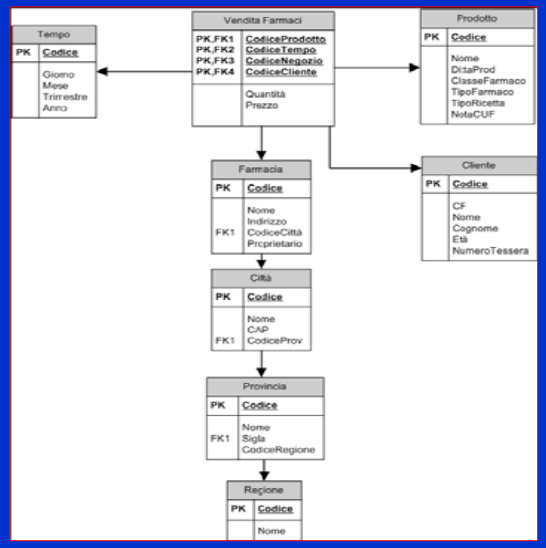
Progettazione logica

- In questa fase si deriva lo schema multi dimensionale individuando ulteriori dimensioni di analisi: il "**tempo**" (in termini di giorno, mese, trimestre e anno) e "**area geografica**" (in termini della città, provincia e regione della farmacia che ha effettuato la vendita) in cui sono state registrate le vendite.

Progettazione fisica

- Nella fase di progettazione fisica si determina lo **schema relazionale a stella** e, per consentire un buon livello di aggregazione delle informazioni, si decide di normalizzare la sola "collocazione geografica" delle farmacie ottenendosi uno schema relazionale a stella del tipo "**a fiocco di neve**".

Lo schema a fiocco di neve



Gestione dei dati

- Su tale DW è poi possibile effettuare in maniera semplice interrogazioni come:
  - selezione del farmaco più venduto in Campania.
  - determinazione dell'età media dei consumatori di un'assegnato farmaco.
  - I clienti di una specifica farmacia.
- Infine vanno "pianificate" apposite procedure di "refreshing" per aggiornare il contenuto del data warehouse ad intervalli di tempo prefissati.