

Data warehousing

Introduzione

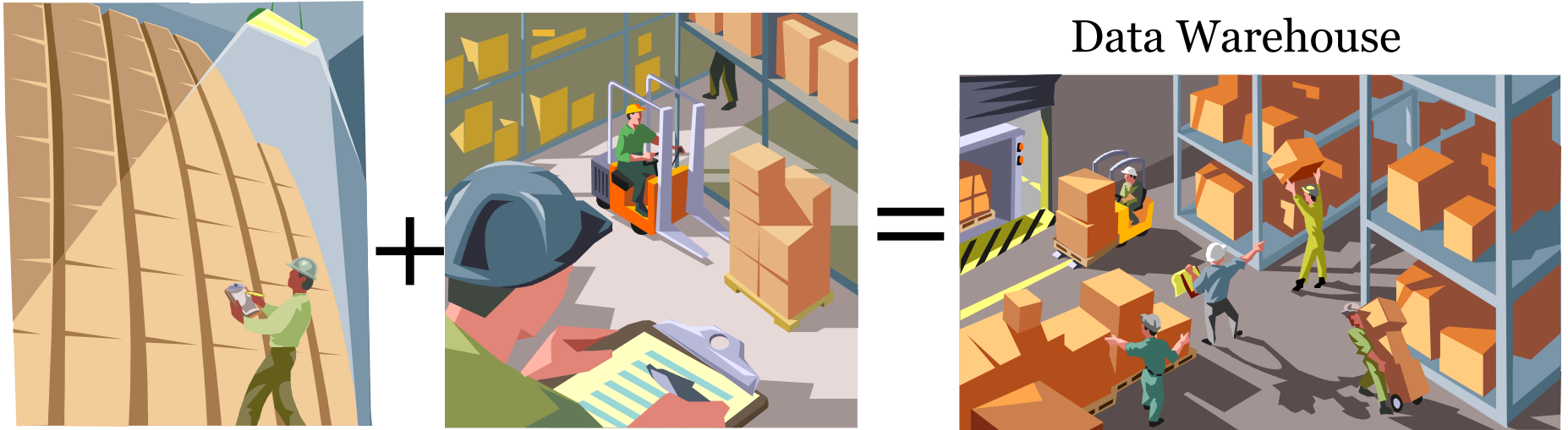
- A partire dagli anni novanta è risultato chiaro che i **database** per i DSS e le analisi di business intelligence vanno **separati** da quelli **operazionali**.
- In questa lezione vedremo le **caratteristiche** di tali database (**data warehouse** e **data mart**), analizzando le differenze rispetto ai sistemi operazionali.
- Analizzeremo gli **aspetti funzionali** dei data warehouse, dando alcuni dettagli relativamente alla loro progettazione.

Un po' di storia

- Nel 1992, William Inmon pubblica il libro *Building the data warehouse*.
- Inmon propone di creare un grande magazzino di dati in cui riversare tutti i dati dell'azienda.
- Dalla aggregazione dei dati è possibile effettuare statistiche su dati aggregati utili ai fini delle analisi.
- Nel 1996, Ralph Kimball propone di eliminare il magazzino e di effettuare le analisi direttamente sulle basi di dati dipartimentali.

Definizione di data warehouse

Magazzino per i dati

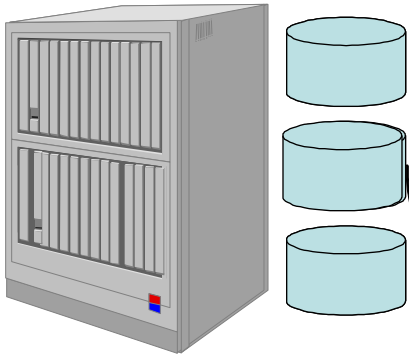


Procedure di acquisizione,
organizzazione e elaborazione

- Data warehousing = l'insieme delle attività per la progettazione, realizzazione ed uso dei data warehouse.

Architettura di data warehouse

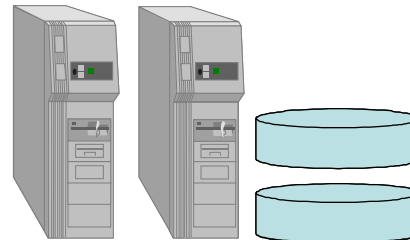
Sistemi Operazionali



Strumenti di Analisi



EnterpriseData Warehouse



Tool OLAP



Fonti Esterne e dati personali



**Tools per Interrogazioni
e Report**

ACQUISIZIONE

INTERROGAZIONI

Motivazione

- Molteplici ragioni inducono a realizzare data warehouse (OLAP) separati dai database transazionali (OLTP):
 - **Integrazione:** i data warehouse integrano dati da fonti diverse
 - **Qualità:** i dati influenzano i risultati
 - **Efficienza:** le analisi devono essere rapide
 - **Estensione temporale:** i dati devono avere sufficiente profondità storica.

Caratteristiche

- Collezione di dati che soddisfa le seguenti proprietà:
 - **Orientata ai soggetti:** considera i dati di interesse ai soggetti dell'organizzazione e non quelli rilevanti ai processi organizzativi
 - **Integrata:** a livello aziendale e non dipartimentale
 - **Storicizzata:** con ampio orizzonte temporale
 - **Consolidata (aggregata):** non interessa “chi” ma “quanti”
 - **Denormalizzata:** le ridondanze permettono tempi di risposta più rapidi.
 - **Fuori linea:** dati aggiornati periodicamente

Differenze OLTP-OLAP

	OLTP	OLAP
Funzione	gestione giornaliera	supporto alle decisioni
Progettazione	orientata alle applicazioni	orientata al soggetto
Frequenza	giornaliera	sporadica
Dati	recenti, dettagliati	storici, riassuntivi, multidim.
Sorgente	singola DB	DB multiple
Uso	ripetitivo	ad hoc
Accesso	read/write	Read
Flessibilità accesso	programmi precompilati	generatori di query
# record acceduti	decine	Migliaia
Utenti	operatori	Manager
# utenti	migliaia	Centinaia
Tipo DB	singola	multiple, eterogenee
Performance	alta	Bassa
Dimensione DB	100 MB - GB	100 GB – TB

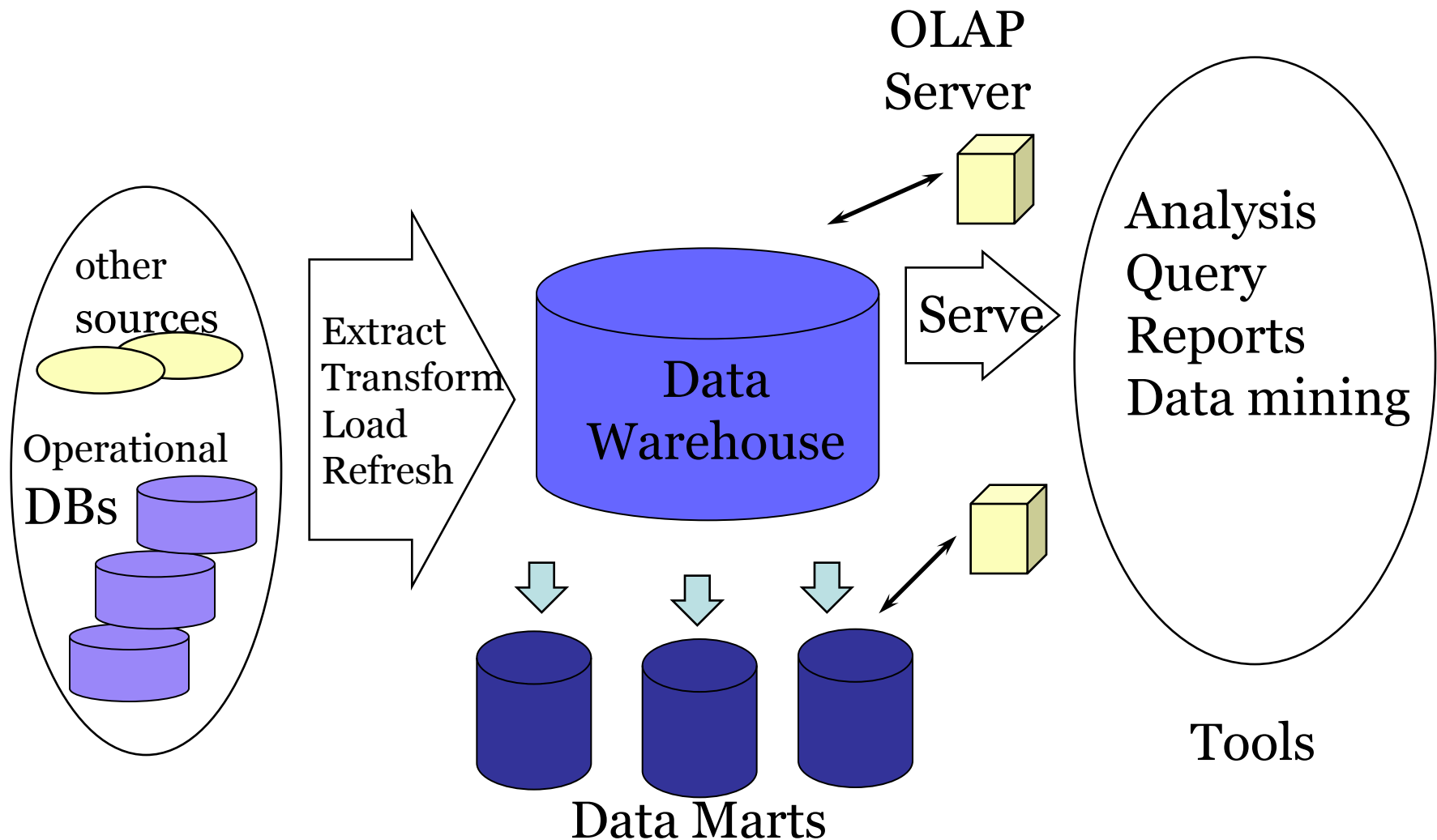
Data mart

- Data warehouse dipartimentale che raccoglie i dati di una specifica funzione aziendale.
- Sistema specializzato che mette insieme i dati necessari ad un dipartimento.
- Implementato creando viste specifiche alle applicazioni.
 - Un data mart di marketing contiene informazioni relative ai clienti e alle transazioni di vendita, risultati di campagne,...
- Sottoinsiemi materializzati di viste dipartimentali che focalizzano su soggetti determinati.

Qualità dei dati

- Verificare, preservare e incrementare la qualità dei dati rappresenta una preoccupazione costante per i responsabili della progettazione e manutenzione.
- Principali inconvenienti:
 - Dati non corretti
 - Dati non aggiornati
 - Dati inesistenti
- I principali fattori che influenzano le analisi di BI riguardano l'accuratezza, completezza, consistenza, attualità, non ridondanza, rilevanza, interpretabilità e accessibilità dei dati.

Architettura del data warehouse



Strumenti ETL

- **Extract, Transform, Load:** insieme degli strumenti che permettono di estrarre, trasformare e caricare i dati
- Nella prima fase i dati vengono estratti dai database operazionali.
 - Estrazione iniziale e incrementale
- Nella fase di trasformazione vengono eliminate le inconsistenze, le duplicazioni, i valori inammissibili.
- I dati corretti e trasformati vengono caricati nel data warehouse.

Realizzazione

- **Top-Down:** Si parte dalle analisi per determinare I dati necessari.
 - Pro: maggiore probabilità di successo,
 - Contro: tempi lunghi.
- **Bottom-up:** Si parte dai dati e si arriva alle analisi
 - Pro: tempi brevi
 - Contro: maggiore probabilità di errore.
- **Ibrida:** Si procede in entrambe le direzioni, realizzando prototipi successivi delle varie parti del sistema
 - Pro: risultati immediati
 - Contro: integrazione incerta.

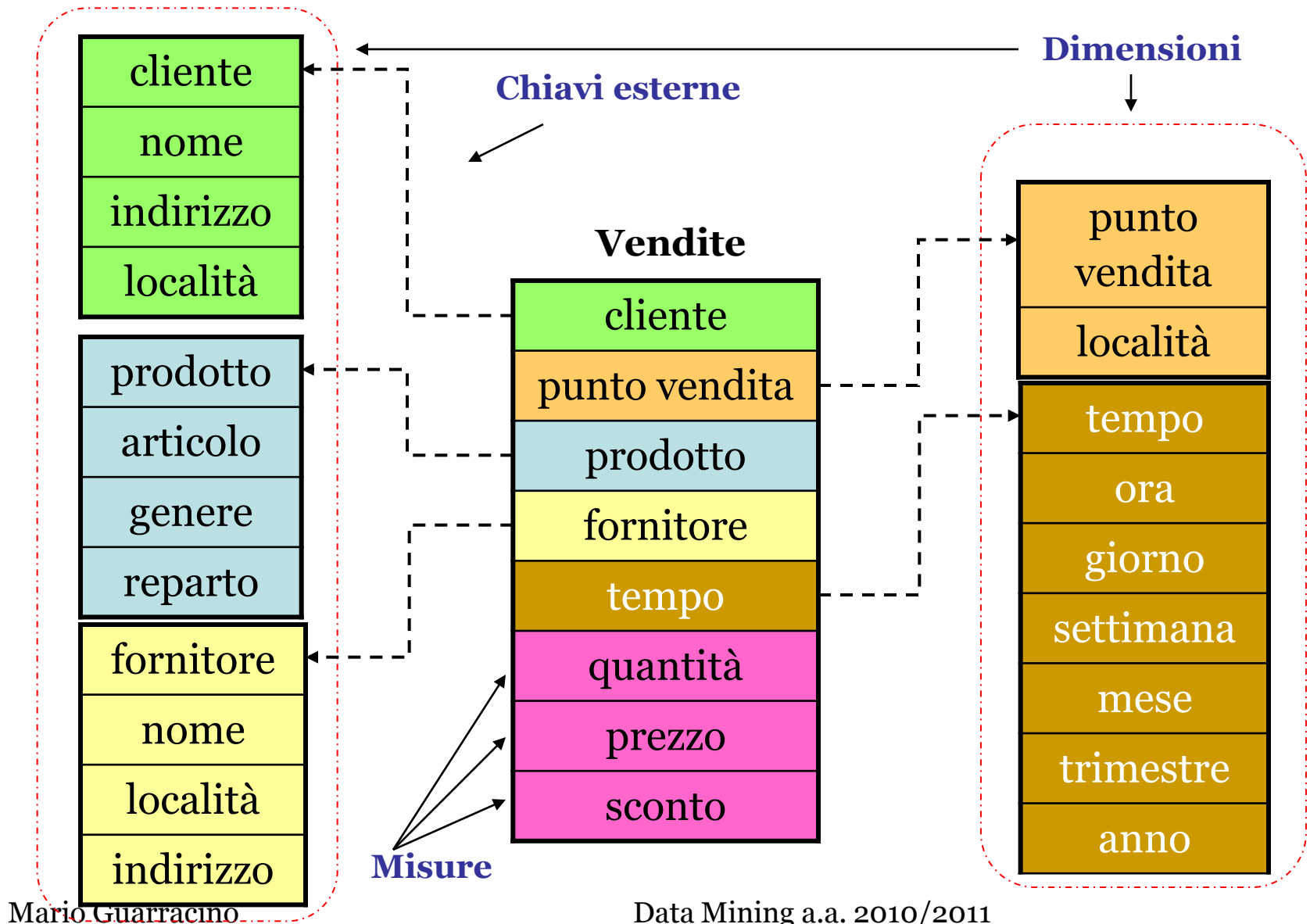
Cubi e analisi multidimensionali

- La modellazione concettuale di un data warehouse utilizza:
- **Schema a stella:** Un singolo oggetto (tabella dei fatti) in mezzo connessa ad un numero di oggetti (tabella delle dimensioni).
- **Schema a fiocco di neve:** Un raffinamento dello schema a stella in cui la gerarchia dimensionale è rappresentata esplicitamente (normalizzando le tabelle delle dimensioni).
- **Galassie:** tabelle dei fatti multiple condividono la tabelle delle dimensioni.

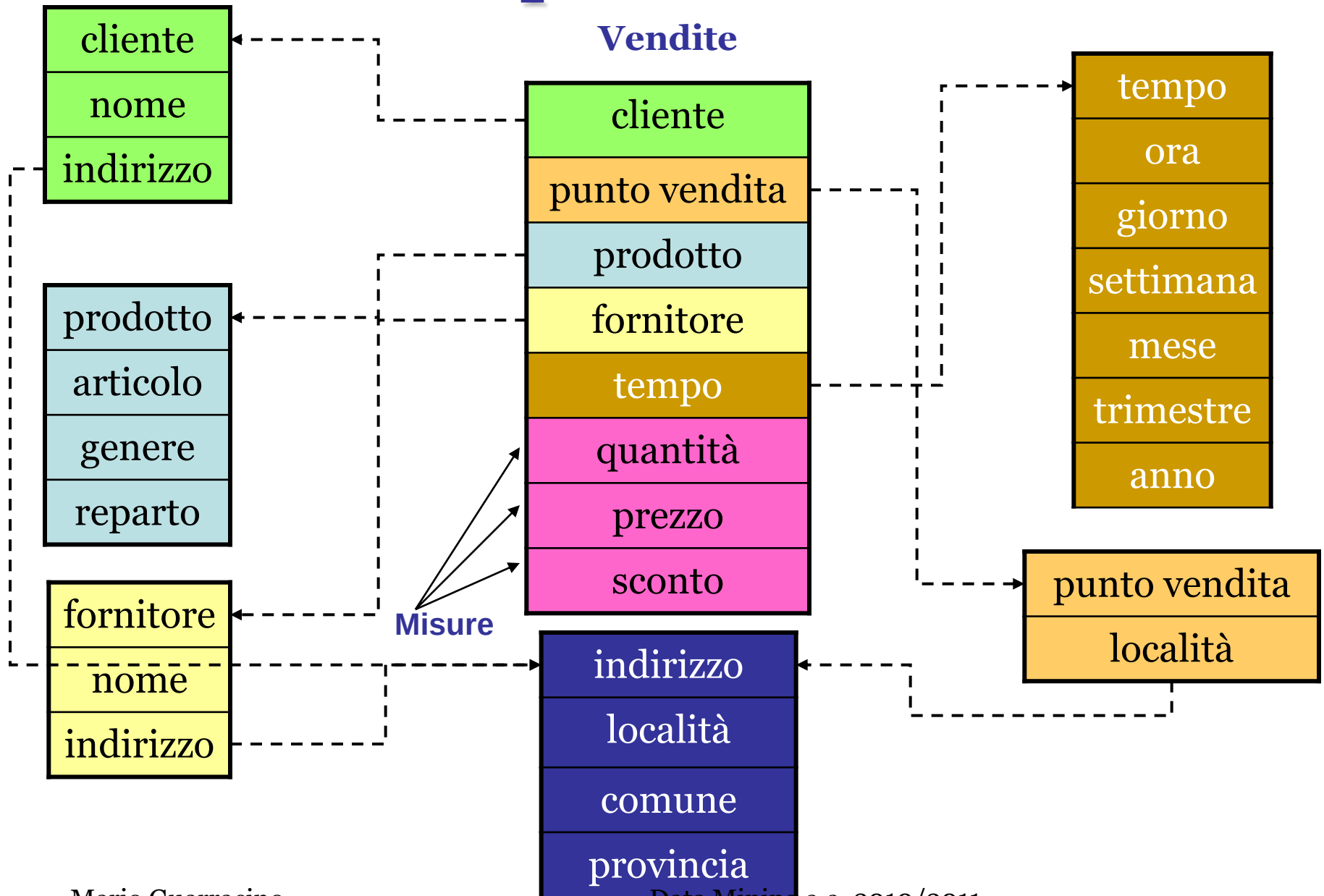
Schema a stella

- Un **fatto** è un **evento** di interesse per l'impresa
 - vendite, spedizioni, acquisti,...
- Le **misure** sono attributi che **descrivono** quantitativamente il **fatto** da diversi punti di vista
 - numero di unità vendute, prezzo unitario, sconto,...
- Una **dimensione** determina la **granularità** minima di **rappresentazione** dei fatti
 - il prodotto, il punto vendita, la data
- Una **gerarchia** determina come le istanze di un fatto possono essere aggregate e selezionate - descrive una dimensione.

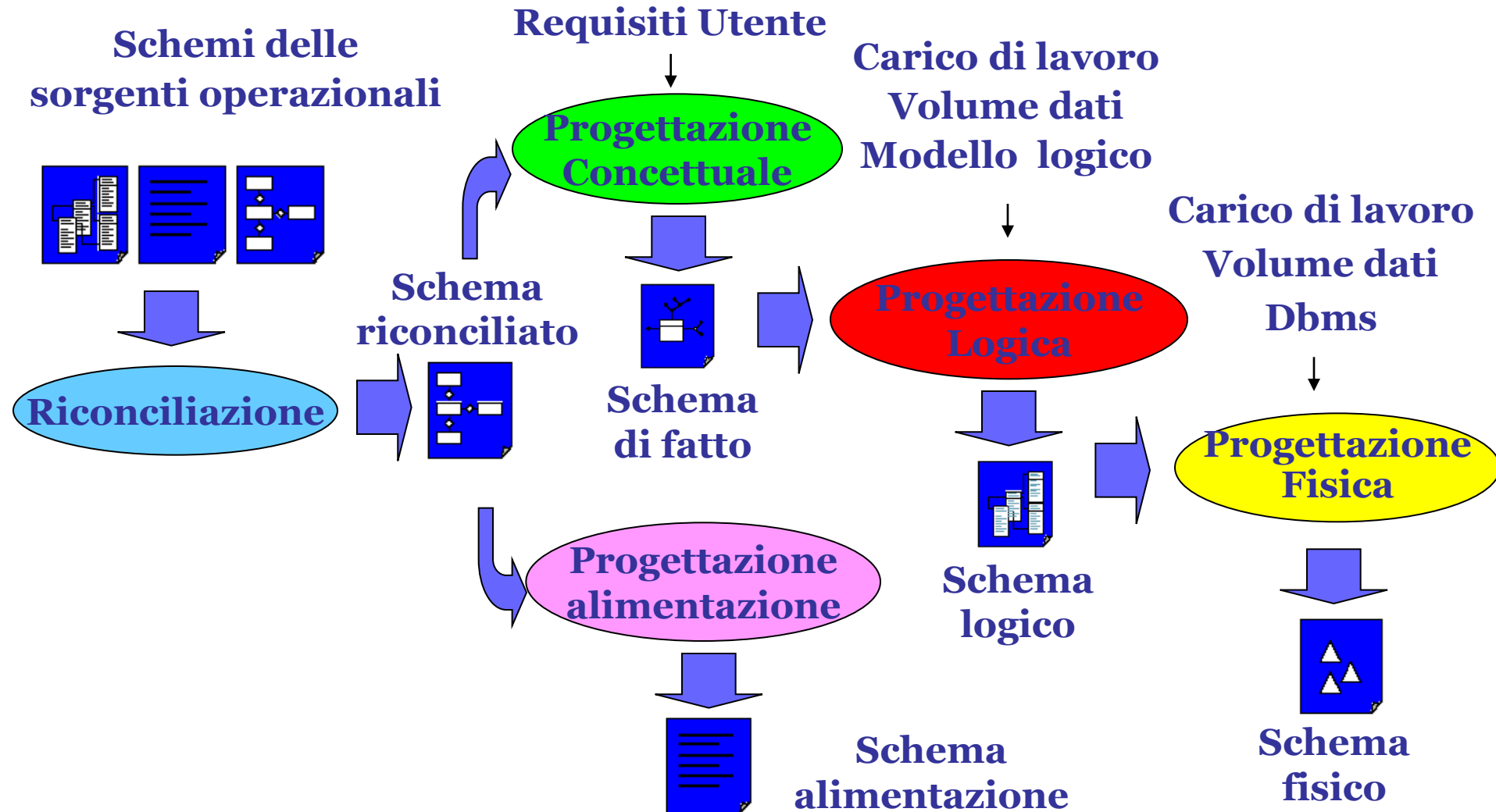
Esempio di schema a stella



Esempio di schema a fiocco



Progettazione dei data mart

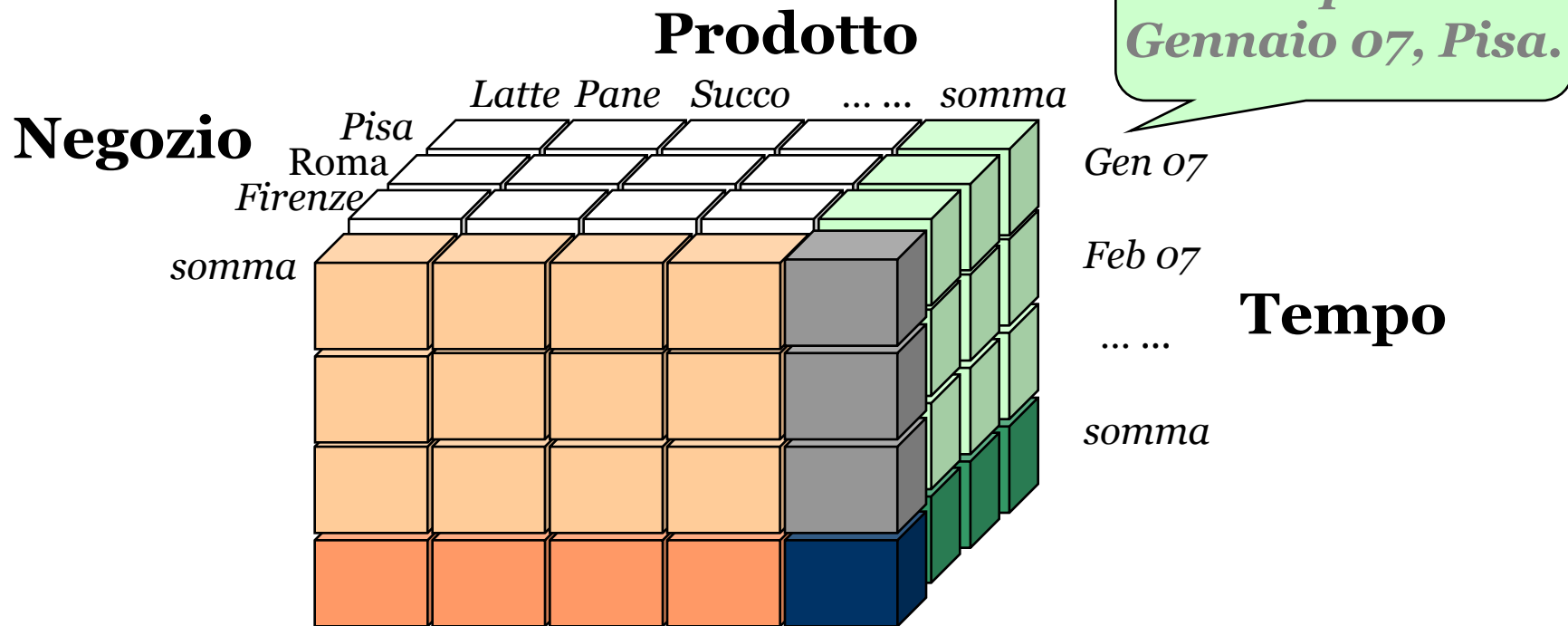


Cubi di dati

- Una tabella dei fatti collegata a n tabelle delle dimensioni può essere rappresentata mediante un cubo di dati a n dimensioni.
- Ogni dimensione contiene una gerarchia di valori e una cella del cubo contiene i valori aggregati
 - count, sum, max, ...
- Essi rappresentano una naturale evoluzione dei fogli elettronici.

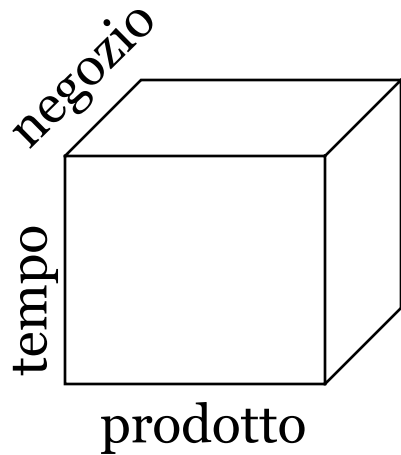
Esempio

- **Tabella dei fatti:** Vendita
- **Dimensioni:** {tempo, prodotto, negozio}
- **Misura:** numero di unità vendute

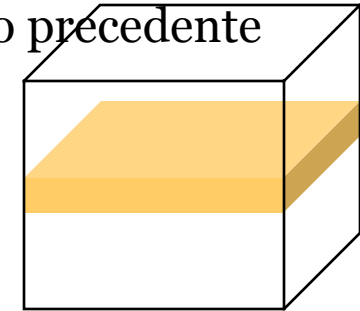


Esempio

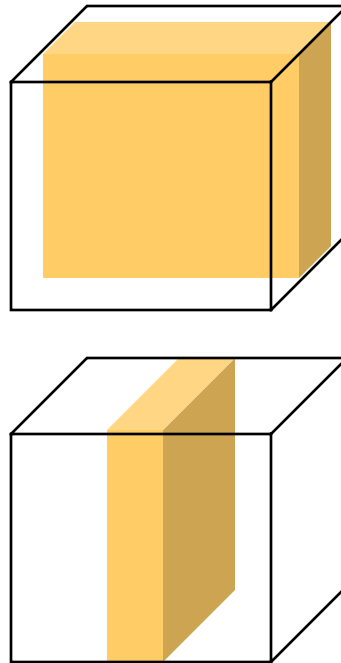
Il manager regionale esamina la vendita di tutti i prodotti in tutti i periodi relativamente ai propri mercati



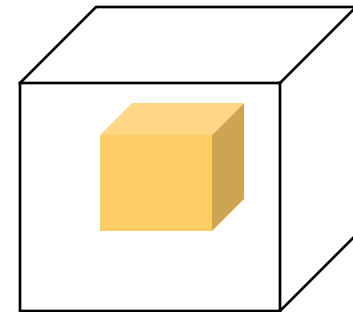
Il manager finanziario esamina la vendita di tutti i prodotti, in tutti i mercati relativamente al periodo corrente e quello precedente



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, un'area regionale e un orizzonte temporale medio



Operazioni sui cubi

- **Roll up**: riassunto dei dati
 - *il volume totale di vendite per categoria di prodotto e per regione*
- **Roll down, drill down, drill through**: passa da un livello di dettaglio basso ad un livello di dettaglio alto
 - *per un particolare prodotto, trova le vendite dettagliate per ogni venditore e per ogni data*
- **Slice and dice**: select & project
 - *Vendite delle bevande nel sud negli ultimi 6 mesi*
- **Pivot**: riorganizza il cubo

Le tabelle Pivot di Excel

- In Excel, le tabelle Pivot consentono di costruire un cubo OLAP a partire da dati memorizzati in una singola tabella
- Le operazioni OLAP corrispondono, in Excel, ad una tecnica di analisi dei dati detta cross-tabulation:
 - a partire da una tabella di dati il comando Pivot Table consente di sintetizzare l'informazione in due (o più modi). Per esempio:
 - vendite raggruppate per prodotto e zona
 - vendite raggruppate per prodotto e mese

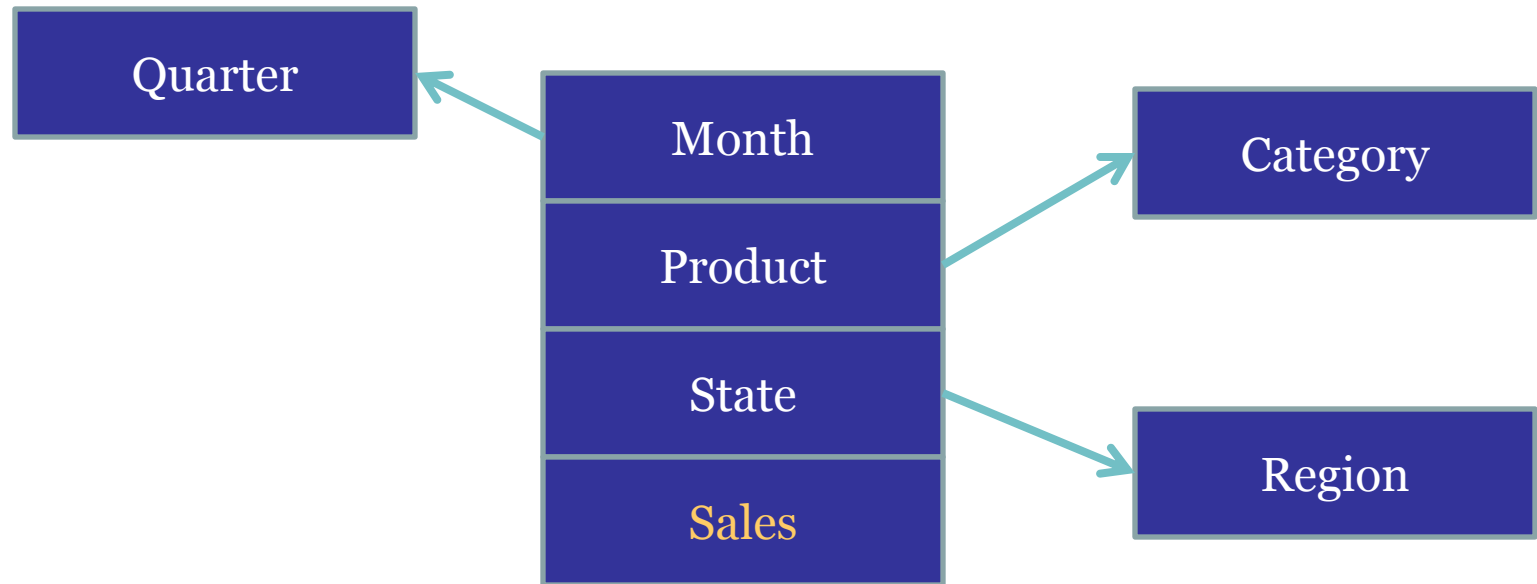
Esempi tratti dal corso *Analisi dei Dati* del prof. Turrini (UniPi)

Creazione delle tabelle pivot

Herbal tea sales.xls [modalità compatibilità] - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Month	Product	State	Sales \$												
2	January	Shining Seas	California	\$644												
3	January	Purple Mountains	California	\$269												
4	January	Purple Mountains	Oregon	\$211												
5	January	Huckleberry Heat	Oregon	\$711												
6	January	Blackbear Berry	California	\$74												
7	January	Raspberry Rocket	California	\$996												
8	January	Blackbear Berry	Oregon	\$211												
9	January	Amber Waves	Oregon	\$45												
10	February	Amber Waves	Washington	\$308												
11	February	Huckleberry Heat	Washington	\$583												
12	February	Purple Mountains	California	\$458												
13	February	Raspberry Rocket	California	\$77												
14	February	Blackbear Berry	California	\$158												
15	February	Shining Seas	California	\$379												
16	February	Amber Waves	Oregon	\$879												
17	February	Huckleberry Heat	Oregon	\$446												
18	February	Purple Mountains	Oregon	\$823												
19	February	Raspberry Rocket	Oregon	\$642												
20	February	Blackbear Berry	Oregon	\$249												
21	February	Shining Seas	Oregon	\$765												
22	February	Amber Waves	Washington	\$120												
23	February	Huckleberry Heat	Washington	\$555												
24	February	Purple Mountains	Washington	\$359												
25	February	Raspberry Rocket	Washington	\$917												
26	February	Blackbear Berry	Washington	\$857												
27	February	Shining Seas	Washington	\$529												
28	March	Amber Waves	California	\$728												
29	March	Huckleberry Heat	California	\$70												
30	March	Purple Mountains	California	\$43												
31	March	Raspberry Rocket	California	\$483												
32	March	Blackbear Berry	California	\$242												

Schema a fiocco



Creazione delle tabelle pivot

The screenshot shows the Microsoft Excel interface with a data table and a 'Crea tabella pivot' (Create PivotTable) dialog box open.

Data Table:

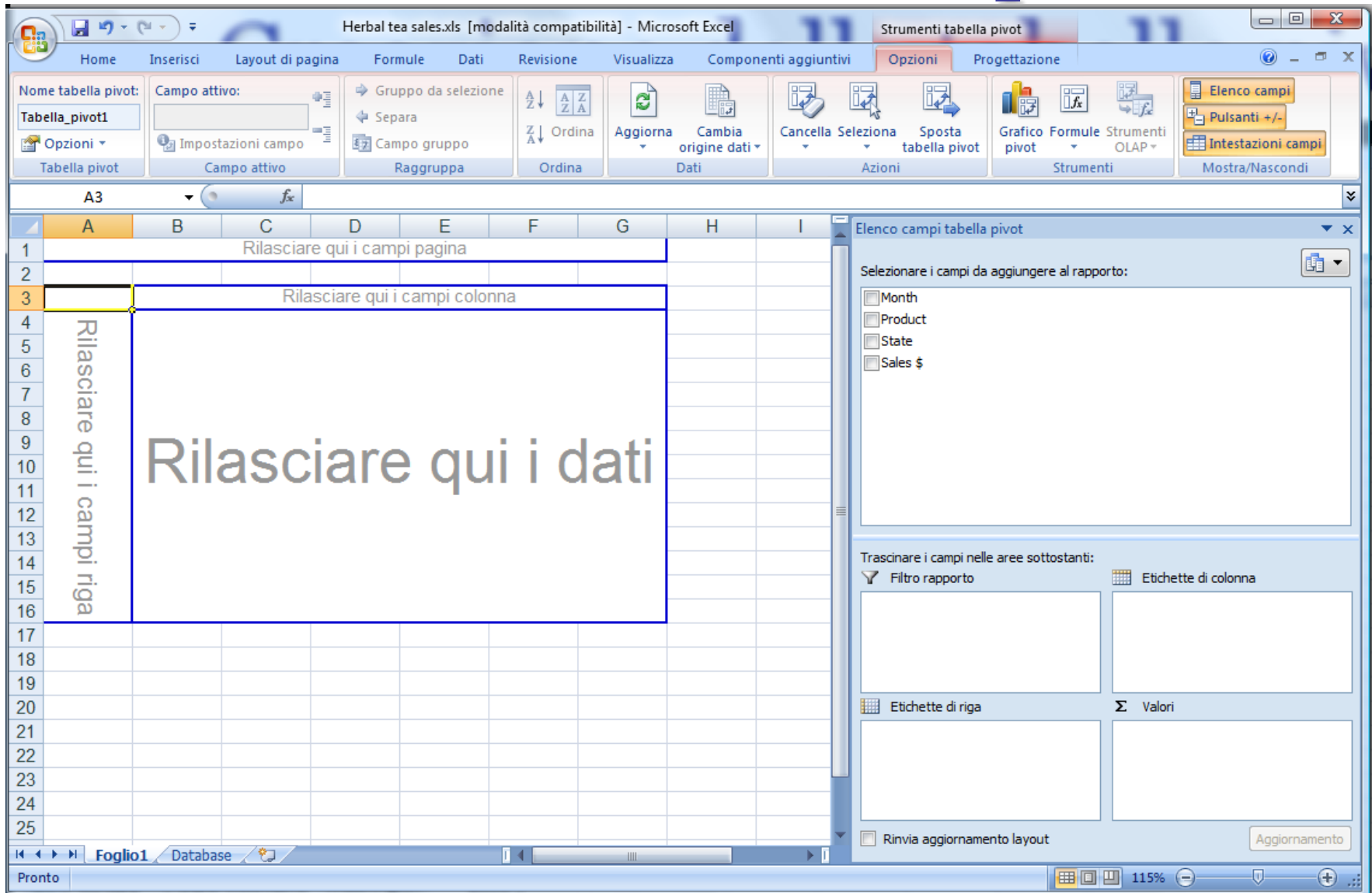
	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	Month	Product	State	Sales \$												
2	January	Shining Seas	California	\$644												
3	January	Purple Mountains	California	\$269												
4	January	Purple Mountains	Oregon	\$211												
5	January	Huckleberry Heat	Oregon	\$711												
6	January	Blackbear Berry	California	\$74												
7	January	Raspberry Rocket	California	\$996												
8	January	Blackbear Berry	Oregon	\$211												
9	January	Amber Waves	Oregon	\$45												
10	February	Amber Waves	Washington	\$308												
11	February	Huckleberry Heat	Washington	\$583												
12	February	Purple Mountains	California	\$458												
13	February	Raspberry Rocket	California	\$77												
14	February	Blackbear Berry	California	\$158												
15	February	Shining Seas	California	\$379												
16	February	Amber Waves	Oregon	\$879												
17	February	Huckleberry Heat	Oregon	\$446												
18	February	Purple Mountains	Oregon	\$823												
19	February	Raspberry Rocket	Oregon	\$642												
20	February	Blackbear Berry	Oregon	\$249												
21	February	Shining Seas	Oregon	\$765												
22	February	Amber Waves	Washington	\$120												
23	February	Huckleberry Heat	Washington	\$555												
24	February	Purple Mountains	Washington	\$359												
25	February	Raspberry Rocket	Washington	\$917												
26	February	Blackbear Berry	Washington	\$857												
27	February	Shining Seas	Washington	\$529												
28	March	Amber Waves	California	\$728												
29	March	Huckleberry Heat	California	\$70												
30	March	Purple Mountains	California	\$43												
31	March	Raspberry Rocket	California	\$483												
32	March	Blackbear Berry	California	\$249												

PivotTable Creation Dialog Box:

- Specificare i dati da analizzare:
 - ☒ Seleziona tabella o intervallo
 - Tabella/Intervallo: Database!\$A\$1:\$D\$225
 - ☐ Origine dati esterna
 - Nome connessione:
- Specificare dove collocare il rapporto di tabella pivot:
 - ☒ Nuovo foglio di lavoro
 - ☐ Foglio di lavoro esistente
 - Posizione:

Buttons: OK, Annulla

Selezione dei campi



Herbal tea sales.xls [modalità compatibilità] - Microsoft Excel

Strumenti tabella pivot

Home Inserisci Layout di pagina Formule Dati Revisione Visualizza Componenti aggiuntivi Opzioni Progettazione

Nome tabella pivot: Tabella_pivot1 Campo attivo: Impostazioni campo

Opzioni Tabella pivot Campo attivo

Gruppo da selezione Separa Campo gruppo Raggruppa

Ordina

Aggiorna Cambia origine dati

Cancella Selezione Sposta tabella pivot Azioni

Grafico pivot Formule Strumenti OLAP

Elenco campi Pulsanti +/- Intestazioni campi Mostra/Nascondi

A3

Rilasciare qui i campi pagina

Rilasciare qui i campi colonna

Rilasciare qui i campi riga

Rilasciare qui i dati

Elenco campi tabella pivot

Selezionare i campi da aggiungere al rapporto:

- ☐ Month
- ☐ Product
- ☐ State
- ☐ Sales \$

Trascinare i campi nelle aree sottostanti:

Filtro rapporto Etichette di colonna

Etichette di riga Valori

Rinvia aggiornamento layout Aggiornamento

Realizzazione della tabella

Herbal tea sales.xls [modalità compatibilità] - Microsoft Excel

Strumenti tabella pivot

Home Inserisci Layout di pagina Formule Dati Revisione Visualizza Componenti aggiuntivi Opzioni Progettazione

Taglia Copia Incolla Copia formato Appunti

Arial 10

Carattere

Testo a capo Allineamento Unisci e centra

Generale Numeri

Formattazione condizionale Formatta come tabella Stili cella

Inserisci Elimina Formato

Celle

Somma automatica Riempimento Cancella Ordina e filtra Trova e seleziona Modifica

A3 Somma di Sales \$

	A	B	C	D	E	F	G	H	I	J	K	L
	Product	April	August	February	January	July	June	March	May	September	Totale complessivo	
5	Amber Waves	1567,815539	1644,09996	1306,510726	45,09634463	1010,038045	861,695487	8530,995464	1510,106643	246,5101198	16722,86833	
6	Blackbear Berry	1640,467959	2034,032423	1263,777749	285,7305774	1410,427182	1529,958555	5814,065279	782,627415	2109,488396	16870,57554	
7	Huckleberry Heat	712,6557782	2635,52905	1584,31371	710,5543577	1812,670217	1212,664329	6833,42738	1004,869387	1168,897639	17675,58185	
8	Purple Mountains	1850,06834	1449,187893	1639,913758	479,6803461	1676,023484	1642,106166	7738,76161	1446,717889	2401,745999	20324,20548	
9	Raspberry Rocket	2060,850446	1957,345067	1635,664274	996,2840728	1126,481172	1357,896841	7629,42204	775,7093319	656,6357749	18196,28902	
10	Shining Seas	1434,14063	2568,293865	1672,764983	643,5241782	917,1312113	1355,725509	8824,262382	1135,900009	1882,180537	20433,9233	
11	Totale complessivo	9265,998693	12288,48826	9102,945199	3160,869877	7952,771311	7960,046887	45370,93416	6655,930675	8465,458466	110223,4435	
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												
22												
23												
24												
25												
26												
27												
28												
29												
30												
31												
32												
33												

Elenco campi tabella pivot

Selezionare i campi da aggiungere al rapporto:

- ☒ Month
- ☒ Product
- ☒ State
- ☒ Sales \$

Trascinare i campi nelle aree sottostanti:

Filtro rapporto: State

Etichette di colore: Month

Etichette di riga: Product

Valori: Somma di Sales \$

Rinvia aggiornamento la... Aggiornamento

Foglio1 Database

Pronto

Operazioni nelle tabelle Pivot

- PIVOT: (ri-)selezionare le variabili da inserire in:
 - Campi riga (asse Y)
 - Campi colonna (asse X)
 - Campi pagina (asse Z)
 - Dati (valori)
- **Roll up**: Raggruppamento di una o più dimensioni in accordo con le gerarchie dello schema a fiocco
- **Drill down**: *Ungrouping* di una o più dimensioni in accordo con le gerarchie stabilite schema a fiocco
- **Slice and dice**: navigazione della tabella Pivot

Esercizio

- Chi ha la maggior probabilità di accedere ad Internet?
- Si consideri l'inchiesta condotta su 1000 utenti Internet scelti in modo casuale, i cui dati sono alla URL

<http://www.di.unipi.it/~turini/Analisi%20dei%20dati/dati/NETUSERS.xls>

- Si costruiscano tabelle Pivot per rispondere alle seguenti domande:
 - Qual è la percentuale di maschi con meno di 30 anni?
 - Quale porzione di questi utenti di Internet corrisponde ai single col solo titolo di scuola superiore?
 - Quale porzione di questi utenti di Internet corrisponde a impiegati?
 - Qual è lo stipendio medio degli impiegati del campione?

Sommario

- Abbiamo visto:
 - Cosa sono i data warehouse;
 - Cosa siano i data mart;
 - Quali siano le architetture dei data warehouse;
 - Cosa siano le tabelle dei fatti, le dimensioni e gli indici;
 - Cosa siano gli schemi a stella e a fiocco;
 - Come si usino i cubi ed analisi multidimensionali;

Nella prossima lezione

- Modelli matematici per le decisioni:
 - Struttura dei modelli di data mining
 - Fasi di sviluppo
 - Classi principali di modelli