



Analisi dei Dati

Lezione 10 – Introduzione al Datawarehouse

Il Datawarehouse

- Il Data Warehousing si può definire come il processo di integrazione di basi di dati indipendenti in un singolo repository (il data warehouse) dal quale gli utenti finali possano facilmente ed efficientemente eseguire query, generare report ed effettuare analisi

Dati Storici

- Correlata alla variabile tempo: presenza di dati storici per eseguire confronti, previsioni e per individuare tendenze
 - basi di dati operazionali: finestra temporale di pochi mesi
 - data warehouse: finestra temporale dell'ordine di anni

Dati Aggregati

- Nell'attività di analisi dei dati per il supporto alle decisioni:
 - non interessa “chi” ma “quanti”
 - non interessa un dato ma la somma, la media, il minimo, il massimo di un insieme di dati

Popolazione di un DW

- Estrazione dei dati dalle sorgenti informative
- Trasformazione e pulizia dei dati, trasformazione di formato e correlazione con oggetti provenienti da altre sorgenti
- Caricamento aggiunta di informazioni temporali e generazione di dati aggregati
- Refresh -modalità incrementale

Schema di un DW

- 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 (num di unità vendute, prezzo unitario)
- Una **dimensione** determina la granularità minima di rappresentazione dei fatti (il prodotto, il negozio, la data)
- Una **gerarchia** determina come le istanze di un fatto possono essere aggregate e selezionate -descrive una dimensione

Dimensioni

- Tipicamente le dimensioni sono:
 - Tempo
 - Collocazione geografica
 - Organizzazione
 - Clienti

Fatti

- I fatti sono tipicamente numerici addittivi
 - Es. vendita in una catena di supermercati i fatti possono essere
 - N. prodotti venduti
 - Incassi
 - Costi
 - ...

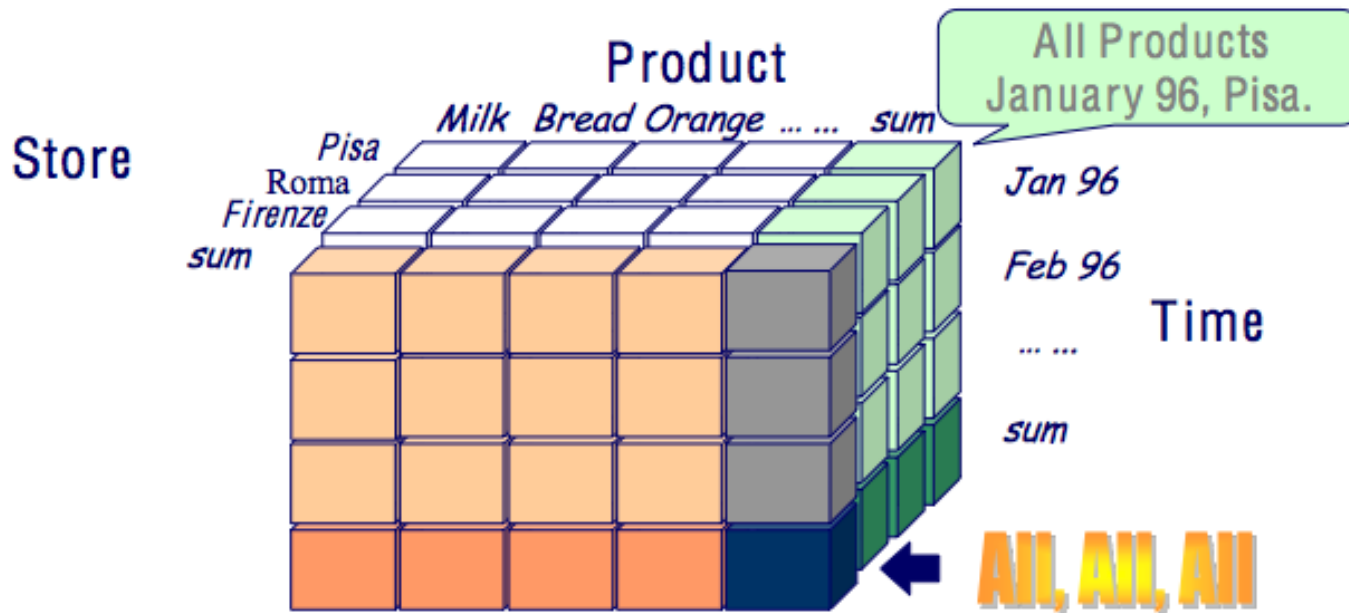
Gerarchie ed Aggregati

- L'idea delle gerarchie é di aggregare automaticamente i dati di interesse quando ci si focalizza su un livello
 - Se ci concentriamo su “Mese” i fatti rappresentano i totali delle vendite per ogni mese
- Possiamo concentrarci su diversi livelli della gerarchia in dimensioni diverse
 - le vendite mensili per regione di ogni prodotto
- Esempio di gerarchia tipica:
 - Comune, Provincia, Regione, Stato, Continente

OLAP: On-Line Analytical Processing

- Una visione multidimensionale, LOGICA, dei dati
- Analisi interattiva dei dati
- Modellazione analitica: derivazione delle proporzioni, delle varianze, etc
- Aggregazioni per ogni intersezione di ogni dimensione.
- Previsione, trend analysis, e statistical analysis.
- Calcola e visualizza i dati in 2D o 3D crosstabs, charts, e grafi, with semplici operazioni di pivoting degli assi

OLAP: Data Cubes



- Ogni dimensione contiene una gerarchia di valori una cella del cubo contiene valori aggregati (count, sum, max, etc.)

OLAP: Operazioni Tipiche

- Roll up: riassumi i 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 West negli ultimi 6 mesi
- Pivot: riorganizza il cubo

Pivot Tables

- Le Pivot Tables di oCalc consentono di costruire un “cubo OLAP” a partire da dati memorizzati in una singola tabella
- Le operazioni OLAP corrispondono, in oCalc, 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
- Un modo più colorito di fare riferimento al trattamento dei dati che si può realizzare con le tabelle Pivot è slice and dice (affettare e ridurre a cubetti) i dati, ovvero organizzare e raggruppare i dati in sottopopolazioni.

Costruzione di pivot tables (I)

- Punto di partenza è una tabella dei dati.
- Esempio: vendita di tisane (HERBALTEAS.xls) sul mercato americano
- Le variabili della tabella sono:
 - mese (month)
 - prodotto (product)
 - stato (state)
 - vendite in dollari (Sales \$)

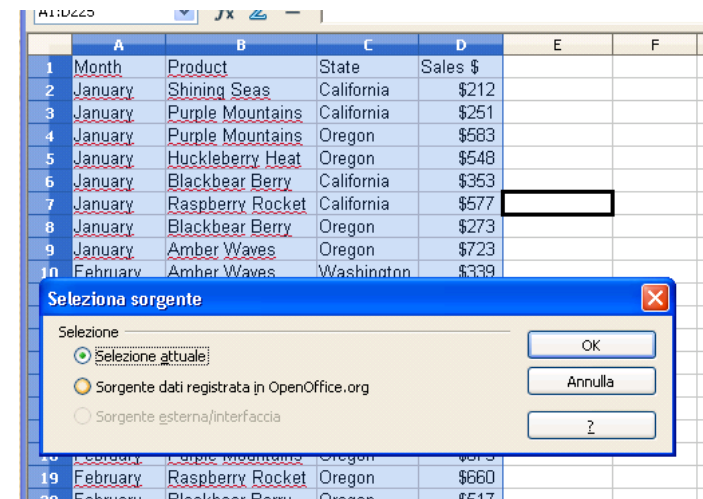
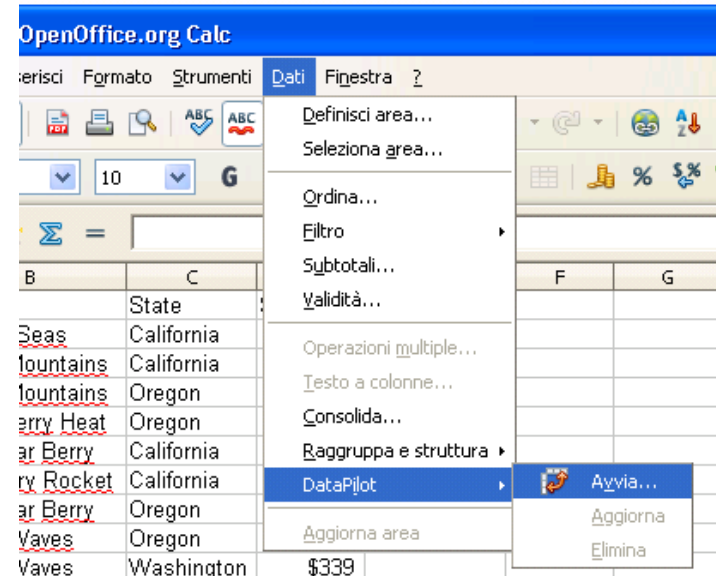
Esempio

The screenshot shows a spreadsheet window titled "Herbal tea sales.xls - OpenOffice.org Calc". The spreadsheet contains data for 17 rows, with columns labeled A through I. The data is as follows:

	A	B	C	D	E	F	G	H	I
1	<u>Month</u>	<u>Product</u>	<u>State</u>	<u>Sales \$</u>					
2	<u>January</u>	<u>Shining Seas</u>	California	\$212					
3	<u>January</u>	<u>Purple Mountains</u>	California	\$251					
4	<u>January</u>	<u>Purple Mountains</u>	Oregon	\$583					
5	<u>January</u>	<u>Huckleberry Heat</u>	Oregon	\$548					
6	<u>January</u>	<u>Blackbear Berry</u>	California	\$353					
7	<u>January</u>	<u>Raspberry Rocket</u>	California	\$577					
8	<u>January</u>	<u>Blackbear Berry</u>	Oregon	\$273					
9	<u>January</u>	<u>Amber Waves</u>	Oregon	\$723					
10	<u>February</u>	<u>Amber Waves</u>	Washington	\$339					
11	<u>February</u>	<u>Huckleberry Heat</u>	Washington	\$352					
12	<u>February</u>	<u>Purple Mountains</u>	California	\$950					
13	<u>February</u>	<u>Raspberry Rocket</u>	California	\$96					
14	<u>February</u>	<u>Blackbear Berry</u>	California	\$780					
15	<u>February</u>	<u>Shining Seas</u>	California	\$24					
16	<u>February</u>	<u>Amber Waves</u>	Oregon	\$928					
17	<u>February</u>	<u>Huckleberry Heat</u>	Oregon	\$249					

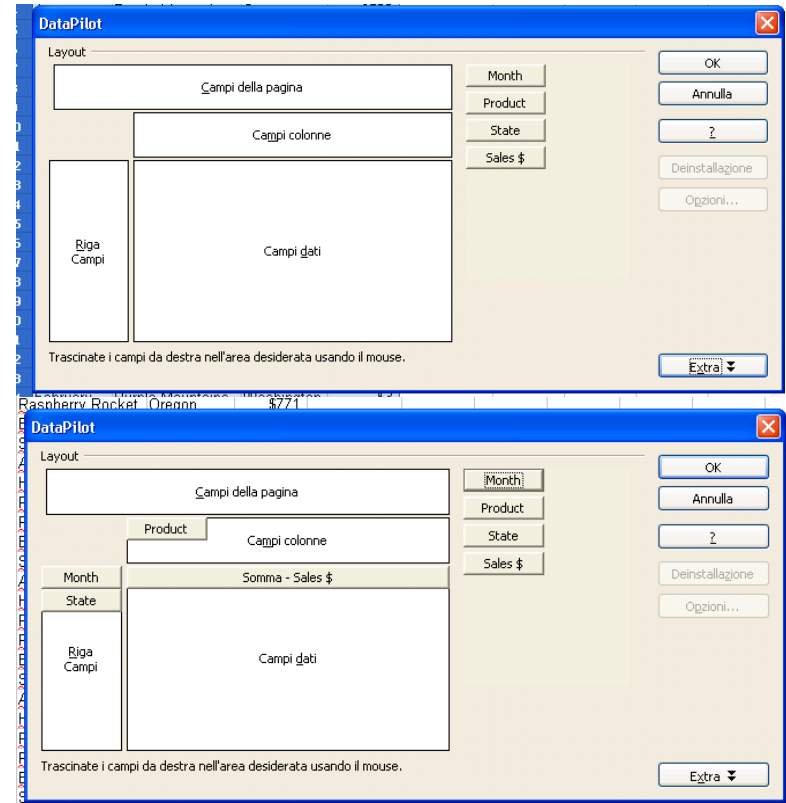
Costruzione di Pivot Tables (2)

- Scegliere Dati → DataPilot → Avvia.. per avviare la creazione guidata
- Rispondere alle domande su dove sono memorizzati i dati da analizzare



Costruzione di Pivot Tables (3)

- Disporre i campi da analizzare negli appositi spazi del layout della tabella pivot
- Premere il bottone “OK”
- In fondo alla tabella dei dati viene inserita la tabella Pivot

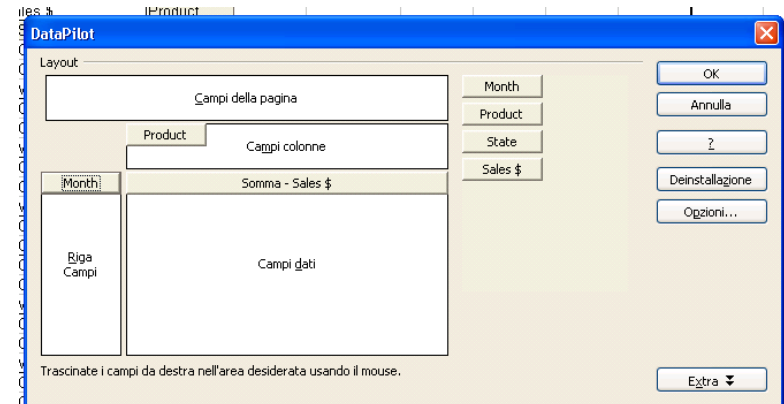


Operazioni OLAP sulle tabelle Pivot

- PIVOT: (ri-)selezionare le variabili da inserire in:
 - ROW FIELDS (asse Y)
 - COLUMN FIELDS (asse X)
 - PAGE FIELDS (asse Z)
 - DATA ITEMS (valori)
- ROLL-UP: Grouping di una o più dimensioni in accordo alle gerarchie stabilite
- DRILL-DOWN: Ungrouping di una o più dimensioni in accordo alle gerarchie stabilite
- Per compiere queste operazioni:
 - Selezionare una cella della tabella pivot e selezionare il menu Dati → DataPilot → Avvia... per ridistribuire i campi da analizzare

Operazioni OLAP: ROLL UP

- Dallo schema della Tabella Pivot rimuovere il campo su cui si vuole aggregare (in questo esempio il campo “State” è stato eliminato)
- La tabella viene aggiornata con la nuova modifica



Month	Product	Totale Risultato
April	Amber Waves	\$985
April	Blackbear	\$1.589
April	Huckleberry	\$1.055
April	Purple Mount	\$2.215
April	Raspberries	\$1.890
April	Shining Seas	\$2.264
April	Totale Risultato	\$9.868
August	Amber Waves	\$1.839
August	Blackbear	\$1.398
August	Huckleberry	\$1.036
August	Purple Mount	\$1.268
August	Raspberries	\$2.364
August	Shining Seas	\$1.728
August	Totale Risultato	\$9.633
February	Amber Waves	\$1.557
February	Blackbear	\$2.469
February	Huckleberry	\$2.189
February	Purple Mount	\$1.297
February	Raspberries	\$2.588
February	Shining Seas	\$1.268
February	Totale Risultato	\$11.368
January	Amber Waves	\$222
January	Blackbear	\$366
January	Huckleberry	\$814
January	Purple Mount	\$1.406
January	Raspberries	\$78
January	Shining Seas	\$869
January	Totale Risultato	\$3.754
July	Amber Waves	\$1.577
July	Blackbear	\$774
July	Huckleberry	\$1.660
July	Purple Mount	\$1.386
July	Raspberries	\$1.544
July	Shining Seas	\$2.326
July	Totale Risultato	\$9.267
June	Amber Waves	\$2.176
June	Blackbear	\$656
June	Huckleberry	\$796
June	Purple Mount	\$1.921
June	Raspberries	\$1.875
June	Shining Seas	\$1.995
June	Totale Risultato	\$9.419
March	Amber Waves	\$9.122
March	Blackbear	\$7.849
March	Huckleberry	\$7.799
March	Purple Mount	\$8.759
March	Raspberries	\$9.284
March	Shining Seas	\$8.171
March	Totale Risultato	\$50.984
May	Amber Waves	\$1.879
May	Blackbear	\$983
May	Huckleberry	\$1.651
May	Purple Mount	\$1.510
May	Raspberries	\$2.510
May	Shining Seas	\$1.025
May	Totale Risultato	\$9.557
September	Amber Waves	\$1.030
September	Blackbear	\$914
September	Huckleberry	\$1.708
September	Purple Mount	\$1.691
September	Raspberries	\$1.246
September	Shining Seas	\$844
September	Totale Risultato	\$7.433
Totale Risultato	Amber Waves	\$20.255
Totale Risultato	Blackbear	\$16.998
Totale Risultato	Huckleberry	\$18.707
Totale Risultato	Purple Mount	\$21.454
Totale Risultato	Raspberries	\$23.379
Totale Risultato	Shining Seas	\$20.488
Totale Risultato	Totale Risultato	\$121.282

Esercitazione

- Svolgimento di un esercizio