

# *Data warehouse*

## *Analisi dei dati*

Elena Baralis

Politecnico di Torino

# Operazioni di analisi dei dati

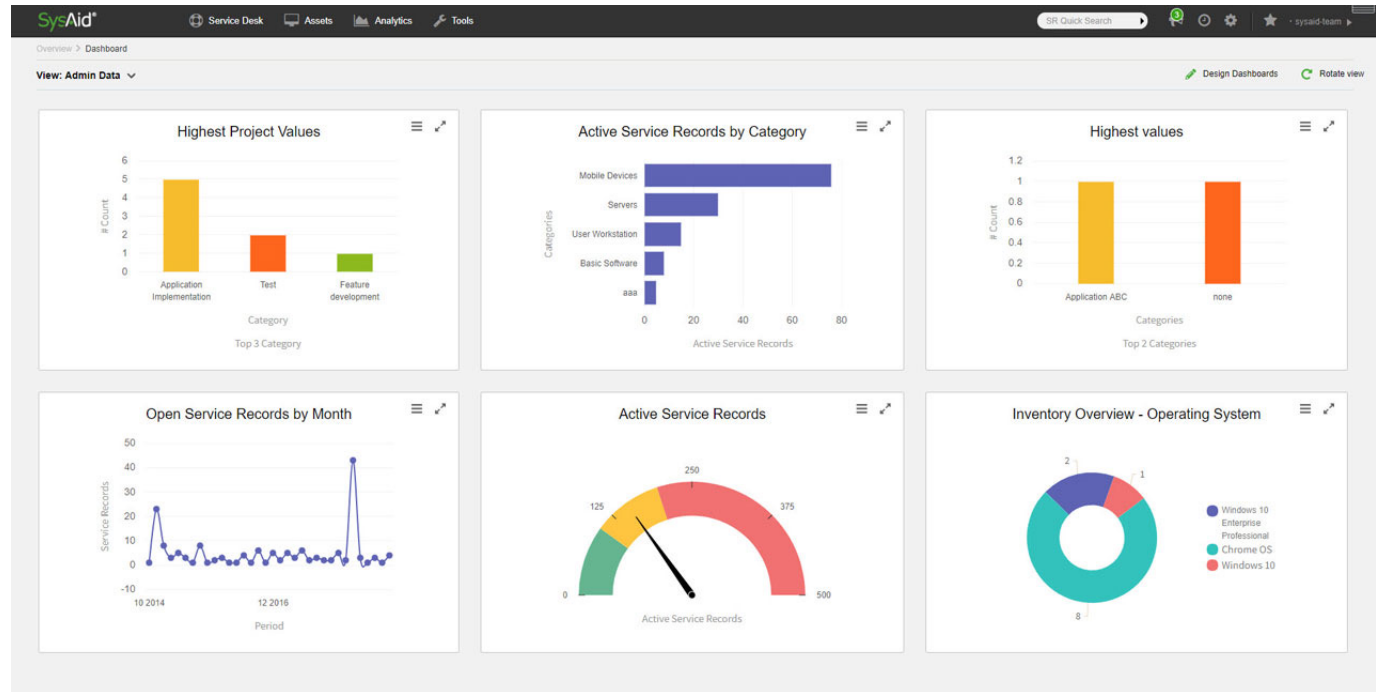
- Calcolo di funzioni aggregate lungo una o più dimensioni
  - necessità di fornire supporto a diversi tipi di funzione aggregata (esempi: media mobile, top ten)
- Operazioni di confronto, essenziali per confrontare l'andamento degli affari (esempio: confronto dei dati delle vendite in mesi diversi)
  - è difficile eseguire confronti utilizzando solo il linguaggio SQL
- Analisi dei dati mediante tecniche di data mining

# Strumenti di interfaccia

L'utente può interrogare il data warehouse mediante strumenti di vario tipo:

- ambiente controllato di query
- strumenti specifici di query e generazione rapporti
- strumenti di data mining

# Ambiente controllato di query



- Sono definite
  - ricerche complesse con struttura prefissata (normalmente parametrica)
  - procedure specifiche di analisi
  - rapporti con struttura prefissata

# Ambiente controllato di query



- È possibile introdurre elementi specifici del settore economico considerato
- È necessario lo sviluppo di codice ad hoc
  - si utilizzano stored procedures, applicazioni contenute in packages, join e aggregazioni predefinite
  - sono disponibili strumenti flessibili per la gestione della reportistica, che permettono di definire layout, periodicità di pubblicazione, liste di distribuzione

# Ambiente di query ad hoc

- È possibile definire interrogazioni OLAP di tipo arbitrario, progettate al momento dall'utente
  - formulazione delle interrogazioni mediante tecniche point and click, che generano automaticamente istruzioni SQL
  - si possono definire interrogazioni (tipicamente) complesse
  - interfaccia basata sul paradigma dello spreadsheet
- Una sessione di lavoro OLAP permette raffinamenti successivi della stessa interrogazione
- Utile quando i rapporti predefiniti non sono adeguati

# *OLAP*

Elena Baralis  
Politecnico di Torino

# Analisi OLAP

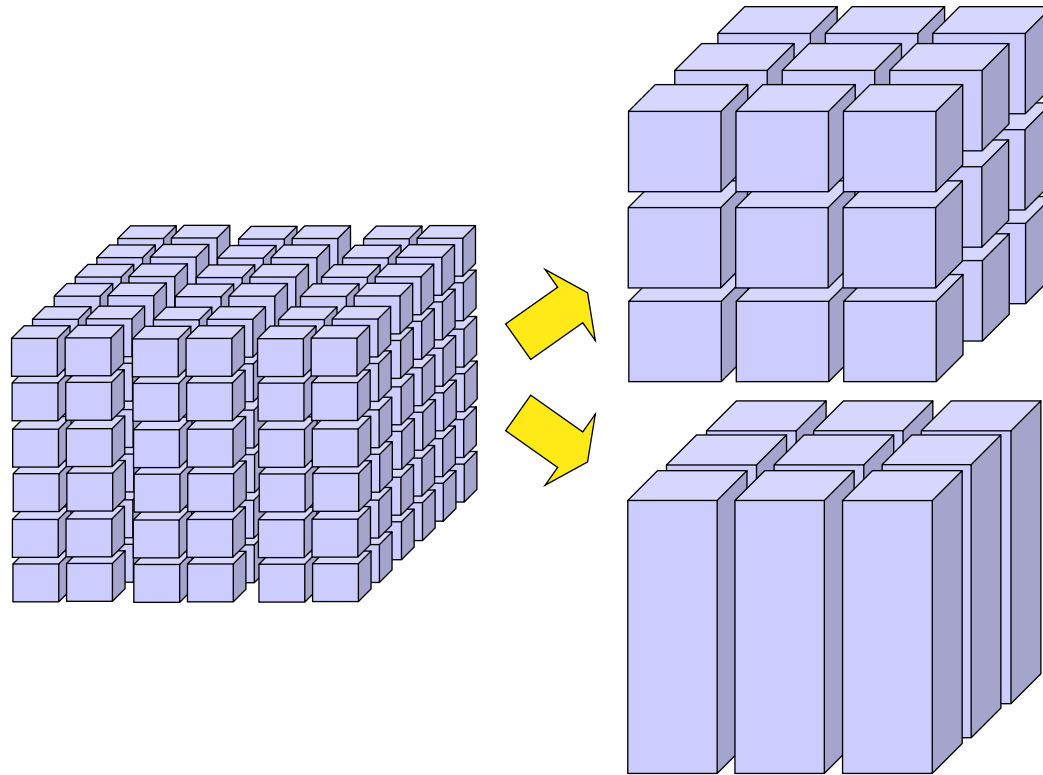
- Operazioni di ricerca disponibili
  - roll up, drill down
  - slice and dice
  - pivot di tabelle
  - ordinamento
- Le operazioni possono essere
  - combinate tra loro nella stessa query
  - eseguite in una sequenza di raffinamenti successivi della stessa query che forma la sessione di lavoro OLAP



# Roll up

- Riduzione di dettaglio dei dati mediante
  - la riduzione del livello di dettaglio di una delle dimensioni presenti, con l'aumento di livello in una gerarchia
    - esempio  
group by negozio, mese → group by città, mese
  - l'eliminazione di una delle dimensioni presenti
    - esempio  
group by prodotto, città → group by prodotto

# Roll up



# Roll up



Metrics Customer Region	Dollar Sales										
	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Month											
Jan 97	\$ 620	\$ 753	\$ 30	\$ 660	\$ 2.405	\$ 1.312	\$ 440	\$ 1.002	\$ 1.002	\$ 383	\$ 210
Feb 97	\$ 258	\$ 252	\$ 800	\$ 975	\$ 160	\$ 582	\$ 744	\$ 310	\$ 799	\$ 118	\$ 357
Mar 97	\$ 648	\$ 244	\$ 148	\$ 250	\$ 1.085	\$ 2.961	\$ 650	\$ 1.240	\$ 119	\$ 142	\$ 96
Apr 97	\$ 787	\$ 588	\$ 447	\$ 486	\$ 226	\$ 506	\$ 601	\$ 119	\$ 550	\$ 85	
May 97	\$ 1.350	\$ 245	\$ 936	\$ 159	\$ 664	\$ 626	\$ 107	\$ 135	\$ 200	\$ 177	\$ 230
Jun 97	\$ 842	\$ 582	\$ 1.281	\$ 937	\$ 240	\$ 774	\$ 176	\$ 1.139	\$ 652	\$ 254	\$ 745
Jul 97	\$ 652	\$ 690	\$ 486	\$ 1.293	\$ 605	\$ 303	\$ 818	\$ 103	\$ 124	\$ 173	\$ 66
Aug 97	\$ 1.783	\$ 304	\$ 1.032	\$ 170	\$ 398	\$ 356	\$ 432	\$ 190	\$ 241	\$ 407	\$ 259
Sep 97	\$ 581	\$ 778	\$ 3.558	\$ 587	\$ 440	\$ 1.652	\$ 1.071	\$ 315	\$ 210	\$ 202	
Oct 97	\$ 2.291	\$ 1.840	\$ 600	\$ 656	\$ 1.300	\$ 718	\$ 1.210	\$ 427	\$ 220	\$ 520	\$ 65
Nov 97	\$ 39	\$ 1.602	\$ 1.082	\$ 1.187	\$ 842	\$ 759	\$ 745	\$ 232	\$ 101	\$ 1.037	\$ 37
Dec 97	\$ 381	\$ 1.588	\$ 343	\$ 118	\$ 1.459	\$ 635	\$ 2.021	\$ 259	\$ 210	\$ 119	\$ 189
Jan 98	\$ 311	\$ 1.174	\$ 2.634	\$ 3.130	\$ 954	\$ 2.083	\$ 1.351	\$ 747	\$ 426	\$ 447	\$ 1.141
Feb 98	\$ 2.518	\$ 702	\$ 1.123	\$ 1.336	\$ 1.227	\$ 3.887	\$ 545	\$ 268	\$ 277	\$ 282	
Mar 98	\$ 2.459	\$ 1.523	\$ 1.178	\$ 4.708	\$ 1.420	\$ 3.514	\$ 1.948	\$ 1.705	\$ 276	\$ 1.168	\$ 63
Apr 98	\$ 407	\$ 841	\$ 524	\$ 712	\$ 133	\$ 2.486	\$ 49	\$ 390	\$ 1.298	\$ 221	\$ 46
May 98	\$ 667	\$ 1.721	\$ 440	\$ 148	\$ 80	\$ 1.310	\$ 303	\$ 104	\$ 657	\$ 65	
Jun 98	\$ 699	\$ 1.096	\$ 898	\$ 353	\$ 902	\$ 839		\$ 230	\$ 155	\$ 105	\$ 75
Jul 98	\$ 586	\$ 1.897	\$ 412	\$ 226	\$ 406	\$ 361	\$ 1.628	\$ 267	\$ 1.011	\$ 41	\$ 184
Aug 98	\$ 894	\$ 326	\$ 792	\$ 1.832	\$ 1.199	\$ 295	\$ 1.816	\$ 277	\$ 102	\$ 118	\$ 115
Sep 98	\$ 338	\$ 3.179	\$ 505	\$ 427	\$ 99	\$ 2.976	\$ 885	\$ 135	\$ 85	\$ 1.110	\$ 510
Oct 98	\$ 544	\$ 413	\$ 1.467	\$ 209	\$ 679	\$ 706	\$ 556	\$ 480	\$ 485	\$ 99	\$ 160
Nov 98	\$ 671	\$ 459	\$ 1.471	\$ 2.066	\$ 701	\$ 716	\$ 986	\$ 1.127	\$ 154	\$ 440	\$ 361
Dec 98	\$ 836	\$ 2.096	\$ 1.726	\$ 3.642	\$ 395	\$ 1.740	\$ 1.943	\$ 1.143	\$ 366	\$ 307	\$ 118



Trimestre, Area

Metrics Customer Region	Dollar Sales										
	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Quarter											
Q1 1997	\$ 1.526	\$ 1.249	\$ 978	\$ 1.885	\$ 3.650	\$ 4.855	\$ 1.834	\$ 2.552	\$ 1.920	\$ 643	\$ 663
Q2 1997	\$ 2.979	\$ 1.415	\$ 2.664	\$ 1.582	\$ 1.130	\$ 1.906	\$ 884	\$ 1.393	\$ 1.402	\$ 516	\$ 975
Q3 1997	\$ 3.016	\$ 1.772	\$ 5.076	\$ 2.050	\$ 1.443	\$ 2.311	\$ 2.321	\$ 608	\$ 575	\$ 782	\$ 325
Q4 1997	\$ 2.711	\$ 5.030	\$ 2.025	\$ 1.961	\$ 3.601	\$ 2.112	\$ 3.976	\$ 918	\$ 531	\$ 1.676	\$ 291
Q1 1998	\$ 5.288	\$ 3.399	\$ 4.935	\$ 9.174	\$ 3.601	\$ 9.484	\$ 3.844	\$ 2.720	\$ 979	\$ 1.897	\$ 1.204
Q2 1998	\$ 1.773	\$ 3.658	\$ 1.862	\$ 1.213	\$ 1.115	\$ 4.635	\$ 352	\$ 724	\$ 2.110	\$ 391	\$ 121
Q3 1998	\$ 1.818	\$ 5.402	\$ 1.709	\$ 2.485	\$ 1.704	\$ 3.632	\$ 4.329	\$ 679	\$ 1.198	\$ 1.269	\$ 809
Q4 1998	\$ 2.051	\$ 2.968	\$ 4.664	\$ 5.917	\$ 1.775	\$ 3.162	\$ 3.485	\$ 2.750	\$ 1.005	\$ 846	\$ 639

# Roll up

Category	Year	Metrics Customer Region	Dollar Sales									
			North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germa
Electronics	1997		\$ 138	\$ 1.774	\$ 384	\$ 138	\$ 2.346	\$ 2.554	\$ 2.184	\$ 566	\$ 199	\$
	1998		\$ 1.184	\$ 4.529	\$ 1.892	\$ 7.232	\$ 651	\$ 9.488	\$ 476	\$ 2.683	\$ 462	\$ 7
Food	1997		\$ 759	\$ 682	\$ 729	\$ 262	\$ 588	\$ 469	\$ 807	\$ 156	\$ 615	\$ 1
	1998		\$ 538	\$ 925	\$ 959	\$ 677	\$ 213	\$ 1.503	\$ 261	\$ 165	\$ 175	\$ 1
Gifts	1997		\$ 2.532	\$ 1.355	\$ 1.854	\$ 1.413	\$ 2.535	\$ 2.132	\$ 1.904	\$ 908	\$ 375	\$ 1.0
	1998		\$ 1.955	\$ 2.785	\$ 2.800	\$ 2.695	\$ 1.813	\$ 2.844	\$ 1.778	\$ 1.158	\$ 717	\$ 6
Health & Beauty	1997		\$ 624	\$ 640	\$ 1.317	\$ 647	\$ 588	\$ 754	\$ 654	\$ 143	\$ 292	\$ 3
	1998		\$ 611	\$ 887	\$ 566	\$ 382	\$ 499	\$ 1.162	\$ 1.044	\$ 273	\$ 72	
Household	1997		\$ 5.354	\$ 4.112	\$ 5.410	\$ 4.446	\$ 3.058	\$ 3.974	\$ 2.654	\$ 3.545	\$ 2.875	\$ 1.9
	1998		\$ 5.787	\$ 5.320	\$ 5.416	\$ 6.812	\$ 4.334	\$ 5.008	\$ 7.588	\$ 2.139	\$ 3.649	\$ 2.7
Kid's Korner	1997		\$ 201	\$ 398	\$ 485	\$ 186	\$ 409	\$ 323	\$ 396	\$ 105	\$ 34	\$
	1998		\$ 247	\$ 422	\$ 441	\$ 380	\$ 221	\$ 592	\$ 290	\$ 198	\$ 19	\$
Travel	1997		\$ 624	\$ 505	\$ 564	\$ 386	\$ 300	\$ 978	\$ 416	\$ 48	\$ 38	
	1998		\$ 608	\$ 559	\$ 1.096	\$ 611	\$ 464	\$ 316	\$ 573	\$ 257	\$ 198	\$



Category	Year	Metrics	Dollar Sales
Electronics	1997		\$ 10.616
	1998		\$ 29.299
Food	1997		\$ 5.300
	1998		\$ 5.638
Gifts	1997		\$ 16.315
	1998		\$ 20.047
Health & Beauty	1997		\$ 6.042
	1998		\$ 5.665
Household	1997		\$ 38.383
	1998		\$ 50.391
Kid's Korner	1997		\$ 2.559
	1998		\$ 2.943
Travel	1997		\$ 4.497
	1998		\$ 4.792

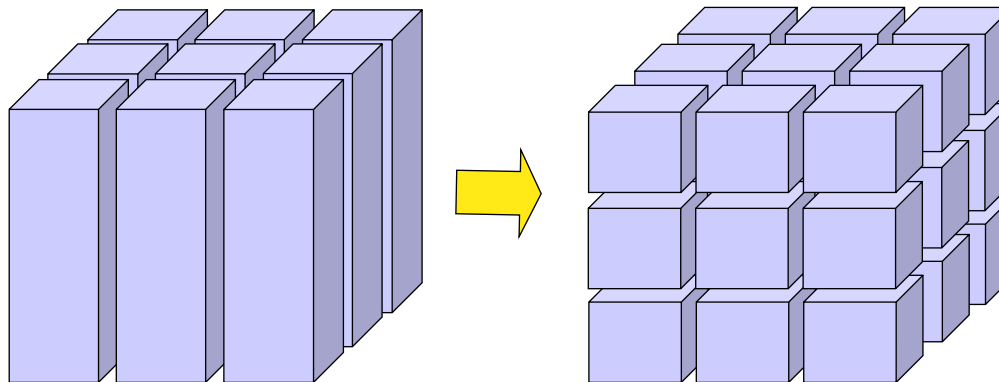
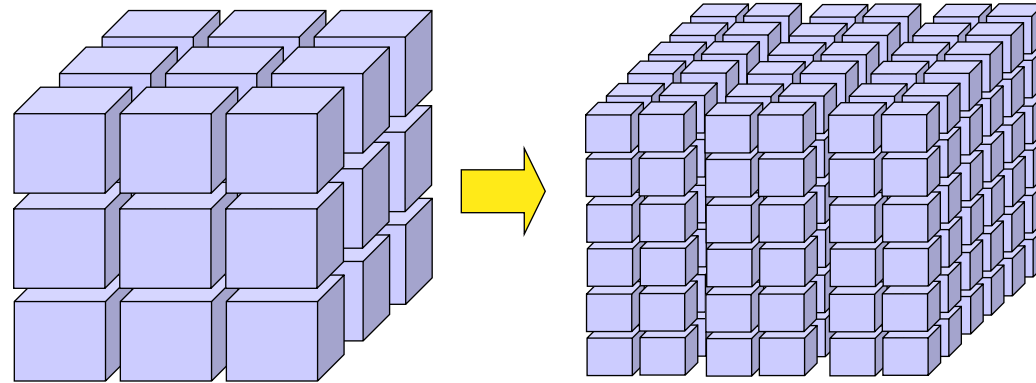
Categoria, Anno

Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

# Drill down

- Aumento di dettaglio dei dati mediante
  - l'aumento del livello di dettaglio di una delle dimensioni presenti, con la riduzione di livello in una gerarchia
    - esempio: da raggruppamento per città e mese a raggruppamento per negozio e mese
  - l'aggiunta di una nuova dimensione
    - esempio: da raggruppamento per città, raggruppamento per città e prodotto
- Spesso il drill down opera su un sottoinsieme dei dati di partenza

# Drill down



# Drill down

Quarter	Metrics Customer Region	Dollar Sales										
		North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Q1 1997		\$ 1.526	\$ 1.249	\$ 978	\$ 1.885	\$ 3.650	\$ 4.855	\$ 1.834	\$ 2.552	\$ 1.920	\$ 643	\$ 663
Q2 1997		\$ 2.979	\$ 1.415	\$ 2.664	\$ 1.582	\$ 1.130	\$ 1.906	\$ 884	\$ 1.393	\$ 1.402	\$ 516	\$ 975
Q3 1997		\$ 3.016	\$ 1.772	\$ 5.076	\$ 2.050	\$ 1.443	\$ 2.311	\$ 2.321	\$ 608	\$ 575	\$ 782	\$ 325
Q4 1997		\$ 2.711	\$ 5.030	\$ 2.025	\$ 1.961	\$ 3.601	\$ 2.112	\$ 3.976	\$ 918	\$ 531	\$ 1.676	\$ 291
Q1 1998		\$ 5.288	\$ 3.399	\$ 4.935	\$ 9.174	\$ 3.601	\$ 9.484	\$ 3.844	\$ 2.720	\$ 979	\$ 1.897	\$ 1.204
Q2 1998		\$ 1.773	\$ 3.658	\$ 1.862	\$ 1.213	\$ 1.115	\$ 4.635	\$ 352	\$ 724	\$ 2.110	\$ 391	\$ 121
Q3 1998		\$ 1.818	\$ 5.402	\$ 1.709	\$ 2.485	\$ 1.704	\$ 3.632	\$ 4.329	\$ 679	\$ 1.198	\$ 1.269	\$ 809
Q4 1998		\$ 2.051	\$ 2.968	\$ 4.664	\$ 5.917	\$ 1.775	\$ 3.162	\$ 3.485	\$ 2.750	\$ 1.005	\$ 846	\$ 639



Quarter	Metrics Customer City	Dollar Sales												
		Arlin	San Pedro	Springfield	Chappel Hill	Scranburg	Pebble Beach	Martinsville	Maddon	Peoria	Pecos	Lake Barkley	Alcameda	Fingers Lake
Q1 1997		\$ 675										\$ 39		
Q2 1997					\$ 203				\$ 53					\$ 135
Q3 1997					\$ 276							\$ 252	\$ 63	
Q4 1997		\$ 215	\$ 124			\$ 113	\$ 45	\$ 192	\$ 348			\$ 79	\$ 98	
Q1 1998				\$ 140	\$ 174			\$ 85			\$ 237	\$ 30	\$ 119	
Q2 1998								\$ 12	\$ 17					
Q3 1998		\$ 734					\$ 25	\$ 1.535						
Q4 1998							\$ 219	\$ 119	\$ 142		\$ 85	\$ 1.533		

Trimestre, Area

# Drill down

Category	Metrics	Dollar Sales	
	Year	1997	1998
Electronics		\$ 10.616	\$ 29.299
Food		\$ 5.300	\$ 5.638
Gifts		\$ 16.315	\$ 20.047
Health & Beauty		\$ 6.042	\$ 5.665
Household		\$ 38.383	\$ 50.391
Kid's Korner		\$ 2.559	\$ 2.943
Travel		\$ 4.497	\$ 4.792

Categoria, Anno



Category	Metrics Customer Region Year	Dollar Sales											
		North-East		Mid-Atlantic		South-East		Central		South		North-West	
		1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998
Electronics		\$ 138	\$ 1.184	\$ 1.774	\$ 4.529	\$ 384	\$ 1.892	\$ 138	\$ 7.232	\$ 2.346	\$ 651	\$ 2.554	\$ 9.488
Food		\$ 759	\$ 538	\$ 682	\$ 925	\$ 729	\$ 959	\$ 262	\$ 677	\$ 588	\$ 213	\$ 469	\$ 1.503
Gifts		\$ 2.532	\$ 1.955	\$ 1.355	\$ 2.785	\$ 1.854	\$ 2.800	\$ 1.413	\$ 2.695	\$ 2.535	\$ 1.813	\$ 2.132	\$ 2.844
Health & Beauty		\$ 624	\$ 611	\$ 640	\$ 887	\$ 1.317	\$ 566	\$ 647	\$ 382	\$ 588	\$ 499	\$ 754	\$ 1.162
Household		\$ 5.354	\$ 5.787	\$ 4.112	\$ 5.320	\$ 5.410	\$ 5.416	\$ 4.446	\$ 6.812	\$ 3.058	\$ 4.334	\$ 3.974	\$ 5.008
Kid's Korner		\$ 201	\$ 247	\$ 398	\$ 422	\$ 485	\$ 441	\$ 186	\$ 380	\$ 409	\$ 221	\$ 323	\$ 592
Travel		\$ 624	\$ 608	\$ 505	\$ 559	\$ 564	\$ 1.096	\$ 386	\$ 611	\$ 300	\$ 464	\$ 978	\$ 316

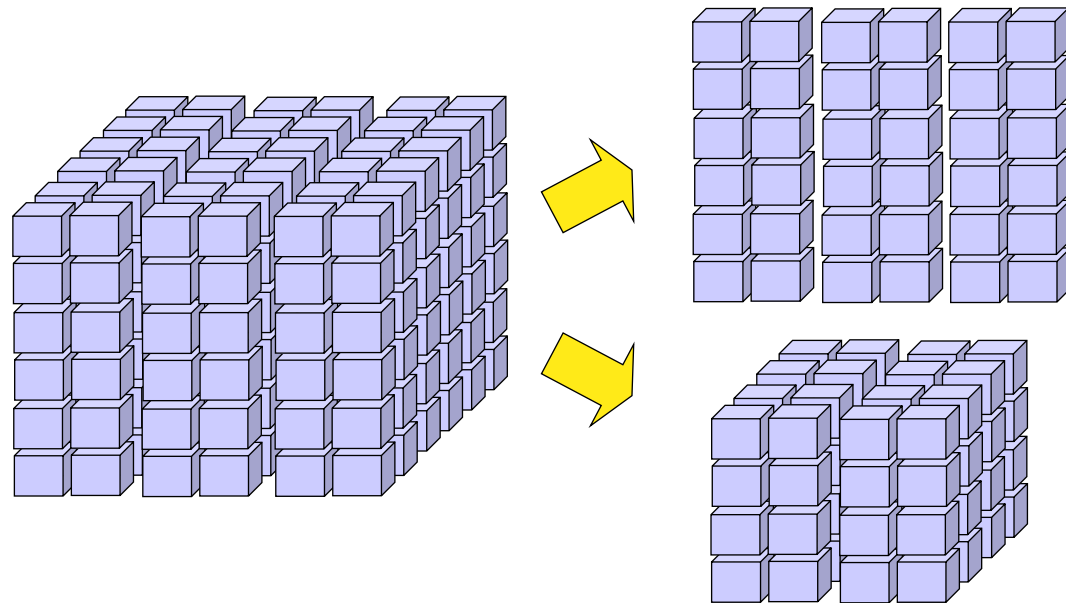
Categoria, Anno, Area



# Slice and dice

- Riduzione del volume dei dati da analizzare
  - selezione di un sottoinsieme mediante predicati
    - slice: predicato di uguaglianza che seleziona una “fetta”
      - esempio: Anno=2005
    - dice: combinazione di predicati che seleziona un “cubetto”
      - esempio: Categoria=‘Alimentari’ and Città=‘Torino’

# Slice and dice



Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

# Slice and dice

Category	Year	Metrics Customer Region	Dollar Sales									
			North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germa
Electronics	1997		\$ 138	\$ 1.774	\$ 384	\$ 138	\$ 2.346	\$ 2.554	\$ 2.184	\$ 566	\$ 199	\$
	1998		\$ 1.184	\$ 4.529	\$ 1.892	\$ 7.232	\$ 651	\$ 9.488	\$ 476	\$ 2.683	\$ 462	\$ 7
Food	1997		\$ 759	\$ 682	\$ 729	\$ 262	\$ 588	\$ 469	\$ 807	\$ 156	\$ 615	\$ 1
	1998		\$ 538	\$ 925	\$ 959	\$ 677	\$ 213	\$ 1.503	\$ 261	\$ 165	\$ 175	\$ 1
Gifts	1997		\$ 2.532	\$ 1.355	\$ 1.854	\$ 1.413	\$ 2.535	\$ 2.132	\$ 1.904	\$ 908	\$ 375	\$ 1.0
	1998		\$ 1.955	\$ 2.785	\$ 2.800	\$ 2.695	\$ 1.813	\$ 2.844	\$ 1.778	\$ 1.158	\$ 717	\$ 6
Health & Beauty	1997		\$ 624	\$ 640	\$ 1.317	\$ 647	\$ 588	\$ 754	\$ 654	\$ 143	\$ 292	\$ 3
	1998		\$ 611	\$ 887	\$ 566	\$ 382	\$ 499	\$ 1.162	\$ 1.044	\$ 273	\$ 72	
Household	1997		\$ 5.354	\$ 4.112	\$ 5.410	\$ 4.446	\$ 3.058	\$ 3.974	\$ 2.654	\$ 3.545	\$ 2.875	\$ 1.9
	1998		\$ 5.787	\$ 5.320	\$ 5.416	\$ 6.812	\$ 4.334	\$ 5.008	\$ 7.588	\$ 2.139	\$ 3.649	\$ 2.7
Kid's Korner	1997		\$ 201	\$ 398	\$ 485	\$ 186	\$ 409	\$ 323	\$ 396	\$ 105	\$ 34	\$
	1998		\$ 247	\$ 422	\$ 441	\$ 380	\$ 221	\$ 592	\$ 290	\$ 198	\$ 19	\$
Travel	1997		\$ 624	\$ 505	\$ 564	\$ 386	\$ 300	\$ 978	\$ 416	\$ 48	\$ 38	
	1998		\$ 608	\$ 559	\$ 1.096	\$ 611	\$ 464	\$ 316	\$ 573	\$ 257	\$ 198	\$



Filter Details:  
Year = 1998

Category	Metrics Customer Region	Dollar Sales										
		North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Ca
Electronics		\$ 1.184	\$ 4.529	\$ 1.892	\$ 7.232	\$ 651	\$ 9.488	\$ 476	\$ 2.683	\$ 462	\$ 702	
Food		\$ 538	\$ 925	\$ 959	\$ 677	\$ 213	\$ 1.503	\$ 261	\$ 165	\$ 175	\$ 100	\$
Gifts		\$ 1.955	\$ 2.785	\$ 2.800	\$ 2.695	\$ 1.813	\$ 2.844	\$ 1.778	\$ 1.158	\$ 717	\$ 686	\$
Health & Beauty		\$ 611	\$ 887	\$ 566	\$ 382	\$ 499	\$ 1.162	\$ 1.044	\$ 273	\$ 72		\$
Household		\$ 5.787	\$ 5.320	\$ 5.416	\$ 6.812	\$ 4.334	\$ 5.008	\$ 7.588	\$ 2.139	\$ 3.649	\$ 2.791	\$
Kid's Korner		\$ 247	\$ 422	\$ 441	\$ 380	\$ 221	\$ 592	\$ 290	\$ 198	\$ 19	\$ 69	
Travel		\$ 608	\$ 559	\$ 1.096	\$ 611	\$ 464	\$ 316	\$ 573	\$ 257	\$ 198	\$ 55	

Anno=1998

# Slice and dice

Subcategory	Metrics Customer City	Dollar Sales												
		Afton	Akron	Albon	Alcameda	Alka	Allagash	Alta	Altoola	Amestra	Amsterdam	Andersonville	Annap	
Audio							\$ 85							
Automotive								\$ 30						
Chocolate		\$ 42	\$ 42		\$ 50		\$ 20		\$ 22	\$ 44			\$	
Christmas		\$ 30					\$ 25	\$ 30	\$ 15					
Classic Toys							\$ 7	\$ 26					\$ 38	
Coffee				\$ 9										
Comfort					\$ 59		\$ 59							
Furniture								\$ 485						
Gadgets								\$ 199	\$ 79	\$ 79				
Games & Puzzles								\$ 17		\$ 45			\$ 45	
Gift Baskets				\$ 55	\$ 43								\$	
Golf		\$ 25							\$ 25	\$ 14			\$ 25	
Hearth										\$ 15				
Jewelry		\$ 75			\$ 189		\$ 24	\$ 77	\$ 189	\$ 24				
Kitchen							\$ 55	\$ 21		\$ 76			\$ :	
Lawn & Garden		\$ 75		\$ 100		\$ 15	\$ 63	\$ 100		\$ 180	\$ 67	\$ 40	\$ :	
Learning		\$ 16							\$ 37					
Meat & Cheese			\$ 40		\$ 20			\$ 20				\$ 25		
Miscellaneous		\$ 200	\$ 1.320			\$ 200	\$ 139			\$ 993				
Natural Remedies		\$ 13								\$ 13				
Pets		\$ 215		\$ 26			\$ 30	\$ 68	\$ 115	\$ 25		\$ 34	\$ :	
Plants & Flowers		\$ 65	\$ 65	\$ 65				\$ 50	\$ 60				\$	
Safety & Security								\$ 30	\$ 22		\$ 22			
Skin Care														
Sleeping				\$ 18										
Toys & Accessories								\$ 29	\$ 185	\$ 744			\$ :	



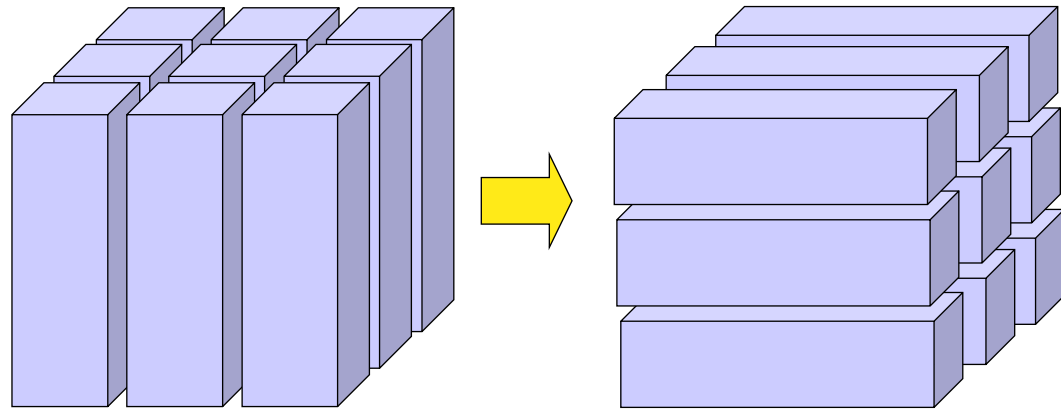
Filter Details:  
 Category = Electronics  
 AND  
 Dollar Sales > 80  
 AND  
 Customer Region = North-West  
 AND  
 Year = 1997

Subcategory	Metrics Customer City	Dollar Sales					
		Alta	Armstrong	Avery Heights	Lane	Mt. Everest	San Francisco
Audio			\$ 98		\$ 123	\$ 85	
Comfort				\$ 118		\$ 1.495	
Gadgets		\$ 199					\$ 199

Tratto da Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

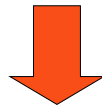
- Riorganizzazione dell'orientamento della struttura multidimensionale senza variare il livello di dettaglio
  - permette una visualizzazione più chiara delle stesse informazioni
  - la rappresentazione dei dati multidimensionali rimane sotto forma di “griglia”
    - due dimensioni sono gli assi principali della griglia
    - varia la posizione delle dimensioni nella griglia

# Pivot



# Pivot

Category	Metrics	Dollar Sales
	Year	
Electronics	1997	\$ 10.616
	1998	\$ 29.299
Food	1997	\$ 5.300
	1998	\$ 5.638
Gifts	1997	\$ 16.315
	1998	\$ 20.047
Health & Beauty	1997	\$ 6.042
	1998	\$ 5.665
Household	1997	\$ 38.383
	1998	\$ 50.391
Kid's Korner	1997	\$ 2.559
	1998	\$ 2.943
Travel	1997	\$ 4.497
	1998	\$ 4.792



Category	Metrics	Dollar Sales	
	Year	1997	1998
Electronics		\$ 10.616	\$ 29.299
Food		\$ 5.300	\$ 5.638
Gifts		\$ 16.315	\$ 20.047
Health & Beauty		\$ 6.042	\$ 5.665
Household		\$ 38.383	\$ 50.391
Kid's Korner		\$ 2.559	\$ 2.943
Travel		\$ 4.497	\$ 4.792

# Pivot

Category	Year	Metrics Customer Region	Dollar Sales									
			North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germa
Electronics	1997		\$ 138	\$ 1.774	\$ 384	\$ 138	\$ 2.346	\$ 2.554	\$ 2.184	\$ 566	\$ 199	\$
	1998		\$ 1.184	\$ 4.529	\$ 1.892	\$ 7.232	\$ 651	\$ 9.488	\$ 476	\$ 2.683	\$ 462	\$ 7
Food	1997		\$ 759	\$ 682	\$ 729	\$ 262	\$ 588	\$ 469	\$ 807	\$ 156	\$ 615	\$ 1
	1998		\$ 538	\$ 925	\$ 959	\$ 677	\$ 213	\$ 1.503	\$ 261	\$ 165	\$ 175	\$ 1
Gifts	1997		\$ 2.532	\$ 1.355	\$ 1.854	\$ 1.413	\$ 2.535	\$ 2.132	\$ 1.904	\$ 908	\$ 375	\$ 1.0
	1998		\$ 1.955	\$ 2.785	\$ 2.800	\$ 2.695	\$ 1.813	\$ 2.844	\$ 1.778	\$ 1.158	\$ 717	\$ 6
Health & Beauty	1997		\$ 624	\$ 640	\$ 1.317	\$ 647	\$ 588	\$ 754	\$ 654	\$ 143	\$ 292	\$ 3
	1998		\$ 611	\$ 887	\$ 566	\$ 382	\$ 499	\$ 1.162	\$ 1.044	\$ 273	\$ 72	
Household	1997		\$ 5.354	\$ 4.112	\$ 5.410	\$ 4.446	\$ 3.058	\$ 3.974	\$ 2.654	\$ 3.545	\$ 2.875	\$ 1.9
	1998		\$ 5.787	\$ 5.320	\$ 5.416	\$ 6.812	\$ 4.334	\$ 5.008	\$ 7.588	\$ 2.139	\$ 3.649	\$ 2.7
Kid's Korner	1997		\$ 201	\$ 398	\$ 485	\$ 186	\$ 409	\$ 323	\$ 396	\$ 105	\$ 34	\$
	1998		\$ 247	\$ 422	\$ 441	\$ 380	\$ 221	\$ 592	\$ 290	\$ 198	\$ 19	\$
Travel	1997		\$ 624	\$ 505	\$ 564	\$ 386	\$ 300	\$ 978	\$ 416	\$ 48	\$ 38	
	1998		\$ 608	\$ 559	\$ 1.096	\$ 611	\$ 464	\$ 316	\$ 573	\$ 257	\$ 198	\$



Category	Metrics Customer Region	Year	Dollar Sales											
			North-East		Mid-Atlantic		South-East		Central		South		North-West	
			1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998
Electronics			\$ 138	\$ 1.184	\$ 1.774	\$ 4.529	\$ 384	\$ 1.892	\$ 138	\$ 7.232	\$ 2.346	\$ 651	\$ 2.554	\$ 9.488
Food			\$ 759	\$ 538	\$ 682	\$ 925	\$ 729	\$ 959	\$ 262	\$ 677	\$ 588	\$ 213	\$ 469	\$ 1.503
Gifts			\$ 2.532	\$ 1.955	\$ 1.355	\$ 2.785	\$ 1.854	\$ 2.800	\$ 1.413	\$ 2.695	\$ 2.535	\$ 1.813	\$ 2.132	\$ 2.844
Health & Beauty			\$ 624	\$ 611	\$ 640	\$ 887	\$ 1.317	\$ 566	\$ 647	\$ 382	\$ 588	\$ 499	\$ 754	\$ 1.162
Household			\$ 5.354	\$ 5.787	\$ 4.112	\$ 5.320	\$ 5.410	\$ 5.416	\$ 4.446	\$ 6.812	\$ 3.058	\$ 4.334	\$ 3.974	\$ 5.008
Kid's Korner			\$ 201	\$ 247	\$ 398	\$ 422	\$ 485	\$ 441	\$ 186	\$ 380	\$ 409	\$ 221	\$ 323	\$ 592
Travel			\$ 624	\$ 608	\$ 505	\$ 559	\$ 564	\$ 1.096	\$ 386	\$ 611	\$ 300	\$ 464	\$ 978	\$ 316



# *Estensioni del linguaggio SQL*

Elena Baralis

Politecnico di Torino

# Estensioni del linguaggio SQL



- Gli strumenti di interfaccia richiedono
  - nuove funzioni aggregate
    - funzioni aggregate utilizzate per le analisi economiche (media mobile, mediana, ...)
    - posizione nell'ordinamento
  - funzioni per la generazione di rapporti
    - definizione di totali parziali e cumulativi
- Lo standard ANSI ha accettato la proposta di nuove funzioni OLAP
  - incorporate nei prodotti a partire da DB2 UDB 7.1, Oracle 8i v2

# Estensioni del linguaggio SQL



- Gli strumenti di interfaccia richiedono
  - operatori per il calcolo di più raggruppamenti (group by) diversi nello stesso momento
- Lo standard SQL-99 (SQL3) ha esteso la clausola group by di SQL

# Base di dati di esempio

Vendite (Città, Mese, Importo)

Città	Mese	Importo
Milano	7	110
Milano	8	10
Milano	9	70
Milano	10	90
Milano	11	35
Milano	12	135
Torino	7	70
Torino	8	35
Torino	9	80
Torino	10	95
Torino	11	50
Torino	12	120

# Funzioni OLAP in SQL

- Nuova classe di funzioni aggregate (funzioni OLAP) caratterizzate da:
  - **finestra di calcolo**, all'interno di cui è possibile specificare il calcolo di funzioni aggregate
    - possibilità di calcolare totali cumulativi e media mobile
  - nuove funzioni aggregate per ricavare la posizione nell'ordinamento (**ranking**)

# Finestra di calcolo

- Nuova clausola **window** caratterizzata da:
  - *partizionamento*: divide le righe in gruppi, senza collassarle (diverso da **group by**)
    - assenza di partizionamento: un solo gruppo
  - *ordinamento delle righe* separatamente all'interno di ogni partizione (simile a **order by**)
  - *finestra di aggregazione*: definisce il gruppo di righe su cui l'aggregato è calcolato, per ciascuna riga della partizione

# Esempio

- Visualizzare, per ogni città e mese
  - l'importo delle vendite
  - la media rispetto al mese corrente e ai due mesi precedenti, separatamente per ogni città

# Esempio

- Partizionamento in base alla città
  - il calcolo della media è azzerato ogni volta che cambia la città
- Ordinamento in base al mese per calcolare la media mobile sul mese corrente insieme ai due mesi precedenti
  - senza ordinamento, il calcolo sarebbe privo di significato
- Dimensione della finestra di calcolo: riga corrente e le due righe precedenti



# Esempio

```
SELECT Città, Mese, Importo,  
       AVG(Importo) OVER Wavg AS MediaMobile  
FROM Vendite  
WINDOW Wavg AS (PARTITION BY Città  
                ORDER BY Mese  
                ROWS 2 PRECEDING)
```

# Esempio

```
SELECT Città, Mese, Importo,  
       AVG(Importo) OVER (PARTITION BY Città  
                          ORDER BY Mese  
                          ROWS 2 PRECEDING)  
       AS MediaMobile  
FROM Vendite
```

# Risultato

Città	Mese	Importo	MediaMobile
Milano	7	110	110
Milano	8	10	60
Milano	9	90	70
Milano	10	80	60
Milano	11	40	60
Milano	12	140	90
Torino	7	70	70
Torino	8	30	50
Torino	9	80	60
Torino	10	100	70
Torino	11	50	60
Torino	12	150	100

Partizione 1

Partizione 2

# Osservazioni

- E` necessario specificare l'ordinamento, perché l'aggregazione richiesta utilizza le righe in modo ordinato
  - l'ordinamento indicato non corrisponde ad un ordine predefinito delle righe in output
- Quando la finestra è incompleta, il calcolo è effettuato sulla parte presente
  - è possibile specificare che, se la finestra è incompleta, il risultato deve essere **NULL**
- E` possibile specificare più finestre di calcolo diverse

# Finestra di aggregazione

- La finestra mobile su cui è effettuato il calcolo dell'aggregato può essere definita
  - a *livello fisico*, formando il gruppo mediante conteggio delle righe
    - esempio: la riga corrente e le due righe precedenti
  - a *livello logico*, formando il gruppo in base alla definizione di un intervallo intorno alla chiave di ordinamento
    - esempio: il mese corrente e i due mesi precedenti

# Definizione intervallo fisico

- Tra un estremo inferiore e la riga corrente

**ROWS 2 PRECEDING**

- Tra un estremo inferiore e uno superiore

**ROWS BETWEEN 1 PRECEDING AND 1 FOLLOWING**

**ROWS BETWEEN 3 PRECEDING AND 1 PRECEDING**

- Tra l'inizio (o la fine) della partizione e la riga corrente

**ROWS UNBOUNDED PRECEDING (o FOLLOWING)**

# Raggruppamento fisico

- Adatto per dati che non hanno interruzioni nella sequenza
  - esempio: non manca nessun mese nella sequenza
  - è possibile specificare più di una chiave di ordinamento
    - il raggruppamento ignora le separazioni
    - esempio: ordinamento per mese e anno
  - non occorrono formule per specificare come calcolare la finestra

# Definizione intervallo logico

- Si utilizza il costrutto **range**, con la stessa sintassi dell'intervallo fisico
- E` necessario definire la distanza tra gli estremi dell'intervallo e il valore corrente sulla chiave di ordinamento
- Esempio

**ORDER BY MONTH**

**RANGE 2 PRECEDING**



# Raggruppamento logico

- Adatto per dati “sparsi”, che hanno interruzioni nella sequenza
  - esempio: manca un mese nella sequenza
  - non è possibile specificare più di una chiave di ordinamento
  - è possibile utilizzare solo tipi di dato numerici o data come chiave di ordinamento (consentono di scrivere espressioni aritmetiche)

# Applicazioni

- Calcolo di aggregati mobili
  - l'aggregato è calcolato su una finestra che “scorre” sui dati
  - esempi: media mobile, somma mobile
- Calcolo di totali cumulativi
  - il totale (cumulativo) è incrementato aggiungendo una riga alla volta
- Confronto tra dati dettagliati e dati complessivi

# Calcolo di totali cumulativi

- Visualizzare, per ogni città e mese
  - l'importo delle vendite
  - l'importo cumulativo delle vendite al trascorrere dei mesi, separatamente per ogni città

# Calcolo di totali cumulativi

- Partizionamento in base alla città
  - il calcolo della somma cumulativa è azzerato ogni volta che cambia la città
- Ordinamento (crescente) in base al mese per calcolare la somma al passare dei mesi
  - senza ordinamento, il calcolo sarebbe privo di significato
- Dimensione della finestra di calcolo: dalla riga iniziale della partizione alla riga corrente

# Calcolo di totali cumulativi



```
SELECT Città, Mese, Importo,  
       SUM(Importo) OVER (PARTITION BY Città  
                          ORDER BY Mese  
                          ROWS UNBOUNDED PRECEDING)  
       AS SommaCumul  
FROM Vendite
```

# Calcolo di totali cumulativi: risultato

Città	Mese	Importo	SommaCumul
Milano	7	110	110
Milano	8	10	120
Milano	9	90	210
Milano	10	80	290
Milano	11	40	330
Milano	12	140	470
Torino	7	70	70
Torino	8	30	100
Torino	9	80	180
Torino	10	100	280
Torino	11	50	330
Torino	12	150	480

Partizione 1

Partizione 2

# Confronto tra dati dettagliati e dati complessivi



- Visualizzare, per ogni città e mese
  - l'importo delle vendite
  - l'importo totale delle vendite sul periodo completo per la città corrente

# Confronto tra dati dettagliati e dati complessivi

- Partizionamento in base alla città
  - il calcolo del totale complessivo è azzerato ogni volta che cambia la città
- Non è necessario l'ordinamento
  - il totale complessivo è calcolato indipendentemente dall'ordinamento
- Non è necessaria la finestra di calcolo
  - è l'intera partizione



# Confronto tra dati dettagliati e dati complessivi



```
SELECT Città, Mese, Importo,  
       SUM(Importo) OVER (PARTITION BY Città)  
       AS ImpTotale  
FROM Vendite
```

# Confronto tra dati dettagliati e dati complessivi

Città	Mese	Importo	ImpTotale
Milano	7	110	470
Milano	8	10	470
Milano	9	90	470
Milano	10	80	470
Milano	11	40	470
Milano	12	140	470
Torino	7	70	480
Torino	8	30	480
Torino	9	80	480
Torino	10	100	480
Torino	11	50	480
Torino	12	150	480

Partizione 1

Partizione 2

# Confronto tra dati dettagliati e dati complessivi

- Visualizzare, per ogni città e mese
  - l'importo
  - il rapporto tra l'importo della riga corrente per le vendite e il totale complessivo
  - il rapporto tra l'importo della riga corrente per le vendite e il totale complessivo per città
  - il rapporto tra l'importo della riga corrente per le vendite e il totale complessivo per mese

# Confronto tra dati dettagliati e dati complessivi

- Tre finestre di calcolo diverse
  - totale complessivo: nessun partizionamento
  - totale per città: partizionamento per città
  - totale per mese: partizionamento per mese
- Non è necessario l'ordinamento per nessuna finestra
  - il totale complessivo è calcolato indipendentemente dall'ordinamento
- La finestra di calcolo è sempre l'intera partizione

# Confronto tra dati dettagliati e dati complessivi



```
SELECT Città, Mese, Importo
       Importo/SUM(Importo) OVER ()
       AS PercTotale
       Importo/SUM(Importo) OVER (PARTITION BY Città)
       AS PercCittà
       Importo/SUM(Importo) OVER (PARTITION BY Mese)
       AS PercMese
FROM Vendite
```

# Confronto tra dati dettagliati e dati complessivi

Città	Mese	Importo	PercTotale	PercCittà	PercMese
Milano	7	110	110/950	110/470	110/180
Milano	8	10	10/950	10/470	10/40
Milano	9	90	90/950	90/470	90/170
Milano	10	80	80/950	80/470	80/180
Milano	11	40	40/950	40/470	40/90
Milano	12	140	140/950	140/470	140/290
Torino	7	70	70/950	70/480	70/180
Torino	8	30	30/950	30/480	30/40
Torino	9	80	80/950	80/480	80/170
Torino	10	100	100/950	100/480	100/180
Torino	11	50	50/950	50/480	50/90
Torino	12	150	150/950	150/480	150/290

# Group by e finestre

- E` possibile abbinare l'uso di finestre con il raggruppamento eseguito dalla clausola **group by**
- La “tabella temporanea” generata dall'esecuzione della clausola **group by** (con eventuale calcolo di funzioni aggregate abbinata al **group by**) diviene l'operando a cui applicare le operazioni definite per la **window**

# Esempio

- Si supponga che la tabella **Vendite** contenga informazioni sulle vendite con granularità giornaliera
- Visualizzare, per ogni città e mese
  - l'importo delle vendite
  - la media rispetto al mese corrente e ai due mesi precedenti, separatamente per ogni città



# Esempio

- E` necessario raggruppare i dati per mese e calcolare l'importo totale per mese prima di effettuare il calcolo della media mobile
  - si usa la clausola group by per calcolare il totale mensile
- La tabella temporanea generata dalla prima aggregazione diviene l'operando su cui definire la finestra di calcolo

# Esempio

```
SELECT Città, Mese, SUM(Importo) AS TotMese,  
        AVG(SUM(Importo)) OVER (PARTITION BY  
        Città  
                                ORDER BY Mese  
                                ROWS 2 PRECEDING)  
        AS MediaMobile  
FROM Vendite, ...  
WHERE <cond. join>  
GROUP BY Città, Mese
```

# Funzioni di ranking

- Funzioni per calcolare la posizione di un valore all'interno di una partizione
  - funzione **rank** () : calcola la posizione, lasciando intervalli vuoti successivi alla presenza di “pari merito”
    - esempio: 2 primi, subito dopo vi è il terzo nella graduatoria
  - funzione **denserank** () : calcola la posizione, senza lasciare intervalli vuoti successivi alla presenza di “pari merito”
    - esempio: 2 primi, subito dopo vi è il secondo nella graduatoria

# Esempio

- Visualizzare, per ogni città nel mese di dicembre
  - l'importo delle vendite
  - la posizione nella graduatoria

# Esempio

- Non occorre partizionamento
  - una sola partizione che include tutte le città
- Ordinamento in base all'importo per stilare la graduatoria
  - senza ordinamento, il calcolo sarebbe privo di significato
- La finestra di calcolo è l'intera partizione

# Esempio

```
SELECT Città, Importo,  
        RANK() OVER (ORDER BY Importo DESC)  
        AS Graduatoria  
FROM Vendite  
WHERE Mese = 12
```

# Risultato

Città	Importo	Graduatoria
Torino	150	1
Milano	140	2

# Ordinamento del risultato

- L'ordinamento del risultato è ottenuto mediante la clausola **order by**
  - può essere diverso dall'ordinamento delle finestre di calcolo
- Esempio: ordinare il risultato dell'esempio precedente in ordine alfabetico di città



# Esempio

```
SELECT Città, Importo,  
       RANK() OVER (ORDER BY Importo DESC)  
       AS Graduatoria  
FROM Vendite  
WHERE Mese = 12  
ORDER BY Città
```

Città	Importo	Graduatoria
Milano	140	2
Torino	150	1

# Estensioni della clausola group by



- Gli spreadsheet multidimensionali richiedono più totali parziali “in un colpo solo”
  - somma delle vendite per mese e città
  - somma delle vendite per mese
  - somma delle vendite per città
- Per motivi di efficienza è opportuno evitare
  - letture multiple dei dati
  - ordinamenti ridondanti dei dati

# Estensioni della clausola `group by`



- Lo standard SQL-99 ha esteso la sintassi della clausola `group by`
  - `rollup` per calcolare le aggregazioni su tutti i gruppi ottenuti togliendo in ordine una colonna per volta dall'insieme specificato di colonne
  - `cube` per calcolare le aggregazioni su tutte le possibili combinazioni delle colonne specificate
  - `grouping sets` per specificare un elenco di raggruppamenti richiesti (diversi da quelli ottenibili con le due clausole precedenti)
    - `()` per richiedere totali generali (nessun raggruppamento)

# Rollup: esempio

- Si considerino le seguenti tabelle
  - Tempo (Tkey, Giorno, Mese, Anno, ...)
  - Supermercato (Skey, Città, Regione, ...)
  - Prodotto (Pkey, NomeP, Marca, ...)
  - Vendite (Skey, Tkey, Pkey, Importo)
- Calcolare il totale delle vendite nel 2000 per le seguenti diverse combinazioni di attributi
  - prodotto, mese e città
  - mese, città
  - città

# Rollup: esempio

```
SELECT Città, Mese, Pkey,  
       SUM(Importo) AS TotVendite  
FROM Tempo T, Supermercato S, Vendite V  
WHERE T.Tkey = V.Tkey  
      AND S.Skey = V.Skey  
      AND Anno = 2000  
GROUP BY ROLLUP (Città, Mese, Pkey)
```

- L'ordinamento delle colonne in `rollup` determina quali aggregati sono calcolati

# Rollup: risultato

Città	Mese	Pkey	TotVendite
Milano	7	145	110
Milano	7	150	10
Milano	...	...	...
Milano	7	<b>NULL</b>	8500
Milano	8	...	...
Milano	<b>NULL</b>	<b>NULL</b>	150000
Torino	...	...	150
Torino	...	<b>NULL</b>	2500
Torino	<b>NULL</b>	<b>NULL</b>	135000
...	...	...	...
<b>NULL</b>	<b>NULL</b>	<b>NULL</b>	25005000

I “superaggregati” sono rappresentati con **NULL**

# Cube: esempio

- Calcolare il totale delle vendite nel 2000 per *tutte* le combinazioni dei seguenti attributi
  - prodotto, mese, città
- Si devono calcolare le seguenti aggregazioni:
  - prodotto, mese, città
  - prodotto, mese
  - mese, città
  - prodotto, città
  - prodotto
  - mese
  - città
  - nessun raggruppamento

# Cube: esempio

```
SELECT Città, Mese, Pkey,  
        SUM(Importo) AS TotVendite  
FROM Tempo T, Supermercato S, Vendite V  
WHERE T.Tkey = V.Tkey  
      AND S.Skey = V.Skey  
      AND Anno = 2000  
GROUP BY CUBE (Città, Mese, Pkey)
```

- L'ordinamento delle colonne in **cube** è ininfluyente



# Calcolo del cubo

- Si considerano le proprietà distributive e algebriche delle funzioni aggregate
  - le funzioni aggregate *distributive* (**min**, **max**, **sum**, **count**) possono essere calcolate a partire da aggregazioni su un numero maggiore di attributi (con granularità maggiore)
    - Esempio: dall'importo totale su prodotto e mese, si calcola l'importo totale per mese
  - per le funzioni aggregate *algebriche* (**avg**, ...) è possibile il calcolo a partire da aggregazioni su un numero maggiore di attributi (con granularità maggiore), pur di memorizzare opportuni risultati intermedi
    - Esempio: per la media serve conoscere
      - il valore della media nel gruppo
      - il numero di elementi per gruppo

# Calcolo del cubo

- Per rendere più efficiente il calcolo del cubo, si usano le proprietà distributive/algebriche delle funzioni aggregate
  - si usano i risultati di **group by** già calcolati
  - l'operazione di **rollup** richiede una sola operazione di ordinamento
  - il cubo può essere visto come una combinazione di più operazioni di **rollup** (in ordine opportuno)
  - si sfruttano operazioni di sort già eseguite (anche parzialmente)
    - è possibile utilizzare l'ordinamento delle colonne (A,B) per ordinare (A,C)

# Grouping Set: esempio



- Calcolare il totale delle vendite nel 2000 per le seguenti combinazioni di attributi
  - mese
  - mese, città, prodotto
- Eseguire un rollup richiederebbe il calcolo di aggregati aggiuntivi

# Grouping Set: esempio



```
SELECT Città, Mese, Pkey,  
        SUM(Importo) AS TotVendite  
FROM Tempo T, Supermercato S, Vendite V  
WHERE T.Tkey = V.Tkey  
      AND S.Skey = V.Skey  
      AND Anno = 2000  
GROUP BY GROUPING SETS  
        (Mese, (Città, Mese, Pkey) )
```