

Data warehouse OLAP

Elena Baralis
Politecnico di Torino

Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 1

*Elena Baralis
Politecnico di Torino*

Data analysis

- OLAP analysis: complex aggregate function computation
 - support to different types of aggregate functions (e.g., moving average, top ten)
- Comparison operations, exploited to compare business trends (example: sale figure comparison for different time periods)
 - difficult by exploiting plain SQL
- Data analysis by means of data mining techniques

Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 2

*Elena Baralis
Politecnico di Torino*

Data analysis tools

- Presentation
 - separate activity: data returned by a query may be rendered by means of different presentation tools
- Motivation search
 - Data exploration by means of progressive, “incremental” refinements (e.g., drill down)

User interface

Users may query the data warehouse by means of various tools:

- controlled query environments
- query and report generation tools
- data mining tools

Controlled query environment

- It encompasses
 - complex queries with predefined structure (usually parametric)
 - ad hoc analysis procedures
 - predefined reports
- Techniques and knowledge of a specific economic area may be exploited
- It requires ad hoc code development
 - stored procedures, application packages, predefined joins and aggregations
 - flexible tools for report management are available, which allow defining
 - report layout
 - publication periodicity
 - distribution list

Ad hoc query environment

- Arbitrary OLAP queries may be defined
- Queries are designed on demand by users
 - query is defined by point and click techniques, which automatically generate SQL instructions
 - (typically) complex queries may be defined
 - spreadsheet is the user interface paradigm
- An OLAP session allows successive refinements of the same query
- Used when predefined reports are not enough

OLAP analysis

- Available query operations
 - roll up, drill down
 - slice and dice
 - (table) pivot
 - sorting
- Operations may be
 - used together in the same query
 - exploited in sequence to refine the same query which builds up the OLAP session

Roll up

- Data detail reduction by
 - decreasing detail in a dimension, by climbing up a hierarchy
 - example
group by store, month → group by city, month
 - dropping a whole dimension
 - example
group by product, city → group by product

Database and data mining group, Politecnico di Torino
DBG

Roll up

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

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 9 Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino
DBG

Roll up

Metrica	Dollar Sales	Customer Region	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Month													
Jan 97	\$ 620	\$ 793	\$ 20	\$ 660	\$ 2495	\$ 1.312	\$ 440	\$ 1.000	\$ 1.000	\$ 383	\$ 218		
Feb 97	\$ 630	\$ 800	\$ 205	\$ 670	\$ 2500	\$ 1.318	\$ 440	\$ 1.010	\$ 798	\$ 318	\$ 227		
Mar 97	\$ 640	\$ 814	\$ 244	\$ 680	\$ 2447	\$ 2.851	\$ 650	\$ 1.020	\$ 119	\$ 350	\$ 235		
Apr 97	\$ 650	\$ 830	\$ 447	\$ 696	\$ 2395	\$ 504	\$ 661	\$ 1.19	\$ 550	\$ 350	\$ 240		
May 97	\$ 2.395	\$ 245	\$ 936	\$ 159	\$ 654	\$ 426	\$ 107	\$ 1.25	\$ 200	\$ 377	\$ 238		
Jun 97	\$ 640	\$ 582	\$ 281	\$ 937	\$ 240	\$ 774	\$ 170	\$ 1.199	\$ 652	\$ 254	\$ 245		
Jul 97	\$ 650	\$ 690	\$ 480	\$ 1.397	\$ 685	\$ 303	\$ 818	\$ 103	\$ 124	\$ 178	\$ 66		
Aug 97	\$ 1.785	\$ 700	\$ 500	\$ 170	\$ 398	\$ 308	\$ 482	\$ 100	\$ 244	\$ 437	\$ 299		
Sep 97	\$ 650	\$ 723	\$ 550	\$ 207	\$ 240	\$ 453	\$ 151	\$ 110	\$ 210	\$ 300	\$ 299		
Oct 97	\$ 2.394	\$ 1843	\$ 680	\$ 656	\$ 2.380	\$ 718	\$ 2.010	\$ 467	\$ 520	\$ 320	\$ 65		
Nov 97	\$ 39	\$ 1.602	\$ 1.080	\$ 1.187	\$ 842	\$ 759	\$ 745	\$ 232	\$ 101	\$ 1.337	\$ 37		
Dec 97	\$ 201	\$ 1.590	\$ 243	\$ 110	\$ 1.459	\$ 625	\$ 2.021	\$ 259	\$ 210	\$ 119	\$ 109		
Jan 98	\$ 311	\$ 1.124	\$ 2.639	\$ 1.130	\$ 984	\$ 2.893	\$ 1.391	\$ 747	\$ 426	\$ 447	\$ 1.143		
Feb 98	\$ 2.818	\$ 702	\$ 3.125	\$ 1.359	\$ 2.827	\$ 3.897	\$ 548	\$ 268	\$ 277	\$ 282			
Mar 98	\$ 650	\$ 520	\$ 4.100	\$ 1.204	\$ 4.100	\$ 2.001	\$ 2.001	\$ 2.001	\$ 2.001	\$ 2.001	\$ 43		
Apr 98	\$ 407	\$ 845	\$ 504	\$ 718	\$ 182	\$ 2.466	\$ 449	\$ 3.900	\$ 2.596	\$ 231	\$ 46		
May 98	\$ 650	\$ 1.725	\$ 440	\$ 140	\$ 80	\$ 1.310	\$ 2.023	\$ 103	\$ 659	\$ 93			
Jun 98	\$ 699	\$ 1.096	\$ 890	\$ 252	\$ 982	\$ 829	\$ 220	\$ 152	\$ 325	\$ 75			
Jul 98	\$ 506	\$ 1.097	\$ 412	\$ 229	\$ 486	\$ 361	\$ 1.628	\$ 267	\$ 3.011	\$ 41	\$ 194		
Aug 98	\$ 694	\$ 328	\$ 792	\$ 1.633	\$ 1.189	\$ 295	\$ 1.618	\$ 277	\$ 182	\$ 318	\$ 115		
Sep 98	\$ 320	\$ 3.179	\$ 4.100	\$ 407	\$ 95	\$ 2.376	\$ 685	\$ 1.195	\$ 1.110	\$ 510			
Oct 98	\$ 649	\$ 1.468	\$ 487	\$ 1.100	\$ 1.100	\$ 1.100	\$ 1.100	\$ 1.100	\$ 1.100	\$ 1.100			
Nov 98	\$ 671	\$ 499	\$ 1.471	\$ 2.086	\$ 781	\$ 716	\$ 980	\$ 1.137	\$ 154	\$ 443	\$ 393		
Dec 98	\$ 226	\$ 2.099	\$ 1.720	\$ 3.642	\$ 295	\$ 1.740	\$ 1.942	\$ 1.142	\$ 266	\$ 397	\$ 118		

Metrica	Dollar Sales	Customer Region	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Quarter													
Q1 1997	\$ 1.526	\$ 1.249	\$ 976	\$ 1.885	\$ 3.850	\$ 4.955	\$ 1.894	\$ 2.550	\$ 1.920	\$ 643	\$ 693		
Q2 1997	\$ 2.809	\$ 1.274	\$ 2.664	\$ 2.582	\$ 1.246	\$ 1.936	\$ 3.984	\$ 1.396	\$ 1.407	\$ 516	\$ 515		
Q3 1997	\$ 3.819	\$ 1.275	\$ 5.177	\$ 2.302	\$ 1.149	\$ 2.117	\$ 2.106	\$ 1.568	\$ 1.579	\$ 579			
Q4 1997	\$ 2.713	\$ 1.200	\$ 3.925	\$ 2.063	\$ 2.801	\$ 2.031	\$ 3.976	\$ 1.510	\$ 521	\$ 1.876	\$ 205		
Q1 1998	\$ 5.256	\$ 3.299	\$ 4.925	\$ 1.174	\$ 2.801	\$ 4.944	\$ 2.044	\$ 2.720	\$ 979	\$ 1.997	\$ 1.204		
Q2 1998	\$ 1.773	\$ 3.650	\$ 1.862	\$ 2.213	\$ 1.315	\$ 4.635	\$ 3.552	\$ 754	\$ 2.110	\$ 391	\$ 1.211		
Q3 1998	\$ 1.818	\$ 5.402	\$ 1.709	\$ 2.485	\$ 1.704	\$ 3.632	\$ 4.329	\$ 679	\$ 1.190	\$ 1.269	\$ 609		
Q4 1998	\$ 2.051	\$ 2.988	\$ 4.864	\$ 9.817	\$ 1.725	\$ 3.682	\$ 3.489	\$ 3.790	\$ 1.005	\$ 946	\$ 639		

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

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 10 Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino
Roll up





Category	Year	Margin										Customer Region	Dollar Sales
		North-East	Mid-Victorian	South-East	Central	South	North-West	South-West	England	France	Germany		
Electronics	1997	\$ 138	\$ 1.774	\$ 384	\$ 139	\$ 2.245	\$ 2.554	\$ 2.184	\$ 266	\$ 199	\$ 1		
	1998	\$ 1.094	\$ 4.529	\$ 1.092	\$ 7.032	\$ 651	\$ 9.400	\$ 475	\$ 2.623	\$ 452	\$ 7		
Food	1997	\$ 799	\$ 682	\$ 729	\$ 262	\$ 581	\$ 469	\$ 807	\$ 356	\$ 375	\$ 1		
	1998	\$ 538	\$ 925	\$ 959	\$ 877	\$ 233	\$ 3.003	\$ 283	\$ 385	\$ 375	\$ 1		
Gifts	1997	\$ 2.932	\$ 1.355	\$ 1.954	\$ 1.419	\$ 2.523	\$ 2.182	\$ 1.484	\$ 968	\$ 375	\$ 1.0		
	1998	\$ 1.821	\$ 1.229	\$ 1.029	\$ 2.895	\$ 1.029	\$ 2.044	\$ 1.779	\$ 1.388	\$ 357	\$ 6		
Health & Beauty	1997	\$ 624	\$ 540	\$ 1.238	\$ 380	\$ 724	\$ 1.044	\$ 1.044	\$ 273	\$ 32	\$ 2		
	1998	\$ 613	\$ 807	\$ 566	\$ 322	\$ 499	\$ 1.162	\$ 1.044	\$ 273	\$ 32	\$ 2		
Household	1997	\$ 5.254	\$ 4.310	\$ 5.438	\$ 4.445	\$ 659	\$ 3.974	\$ 654	\$ 545	\$ 2.875	\$ 1.9		
	1998	\$ 5.707	\$ 5.320	\$ 5.435	\$ 6.812	\$ 4.334	\$ 5.000	\$ 7.588	\$ 2.139	\$ 3.649	\$ 2.7		
Kid's corner	1997	\$ 201	\$ 398	\$ 485	\$ 395	\$ 493	\$ 323	\$ 398	\$ 105	\$ 34	\$		
	1998	\$ 247	\$ 422	\$ 445	\$ 383	\$ 225	\$ 992	\$ 298	\$ 398	\$ 39	\$		
Total	1997	\$ 624	\$ 905	\$ 1.064	\$ 395	\$ 389	\$ 976	\$ 418	\$ 48	\$ 35	\$		
	1998	\$ 928	\$ 559	\$ 1.095	\$ 811	\$ 494	\$ 216	\$ 573	\$ 257	\$ 280	\$		

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

Elena Baralis
Politecnico di Torino

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 11

Database and data mining group, Politecnico di Torino
Drill down



- Data detail increase by
 - increasing detail in a dimension, by walking down a hierarchy
 - example
group by city, month → group by store, month
 - adding a whole dimension
 - example
group by product → group by product, city
- Frequently drill down operates on a subset of data produced by the initial query

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 12 **Elena Baralis**
Politecnico di Torino

Database and data mining group, Politecnico di Torino
DBG

Drill down

From Galfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 13 Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino
DBG

Drill down

Quarter	Customer Region	Dollar Sales
Q1 1997	North-East	\$ 1.526
Q2 1997	Mid-Atlantic	\$ 2.490
Q3 1997	South-East	\$ 2.876
Q4 1997	Central	\$ 2.850
Q1 1998	South	\$ 2.885
Q2 1998	North-West	\$ 2.850
Q3 1998	South-West	\$ 2.850
Q4 1998	England	\$ 2.850
Q1 1999	France	\$ 2.850
Q2 1999	Germany	\$ 2.850
Q3 1999	Canada	\$ 2.850
Q4 1999		\$ 2.850

Quarter	Customer Cts	Dollar Sales
Q1 1997	Retail	\$ 675
Q2 1997	Sam's Club	\$ 203
Q3 1997	Banfield	\$ 276
Q4 1997	Chappel Hill	\$ 113
Q1 1998	Scranton	\$ 45
Q2 1998	Pebble Beach	\$ 192
Q3 1998	Martinsville	\$ 348
Q4 1998	Maddon	\$ 53
Q1 1999	Peoria	\$ 39
Q2 1999	Peoria	\$ 129
Q3 1999	Lake	\$ 292
Q4 1999	Battier	\$ 63
Q1 2000	Lakeside	\$ 79
Q2 2000	Rogers	\$ 237
Q3 2000	Lake	\$ 30
Q4 2000	Rogers	\$ 119

From Galfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 14 Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino
Drill down

The diagram illustrates a 'drill down' process. At the top is a detailed sales table for the year 1998, categorized by product type (Electronics, Food, Gifts, Household & Retail, Household, Kid's Corner, Travel) and year (1997, 1998). An orange arrow points downwards to a second table, which shows sales data grouped by Customer Region (North-East, Mid-West, South-East, Central, South, North-West) and year (1997, 1998).

Category	Metrics		Dollar Sales		Year	
	1997	1998	\$ 19.616	\$ 19.546	1997	1998
Electronics	\$ 10.616	\$ 10.546	\$ 1.774	\$ 1.774	\$ 1.774	\$ 1.774
Food	\$ 7.599	\$ 7.536	\$ 4.529	\$ 4.529	\$ 4.529	\$ 4.529
Gifts	\$ 2.832	\$ 2.779	\$ 2.115	\$ 2.115	\$ 2.115	\$ 2.115
Household & Retail	\$ 8.042	\$ 8.042	\$ 2.315	\$ 2.047	\$ 2.315	\$ 2.047
Household	\$ 38.283	\$ 38.283	\$ 9.568	\$ 9.568	\$ 9.568	\$ 9.568
Kid's Corner	\$ 2.559	\$ 2.543	\$ 2.559	\$ 2.543	\$ 2.559	\$ 2.543
Travel	\$ 4.487	\$ 4.392	\$ 4.487	\$ 4.392	\$ 4.487	\$ 4.392

Customer Region	Metrics		Dollar Sales		Year	
	1997	1998	1997	1998	1997	1998
North-East	\$ 130	\$ 1104	\$ 1.774	\$ 1.774	\$ 1.774	\$ 1.774
Mid-West	\$ 759	\$ 530	\$ 602	\$ 625	\$ 729	\$ 759
South-East	\$ 2.832	\$ 1.959	\$ 1.799	\$ 2.115	\$ 2.800	\$ 2.453
Central	\$ 8.042	\$ 8.042	\$ 2.315	\$ 2.047	\$ 2.047	\$ 2.047
South	\$ 38.283	\$ 38.283	\$ 9.568	\$ 9.568	\$ 9.568	\$ 9.568
North-West	\$ 2.559	\$ 2.543	\$ 2.559	\$ 2.543	\$ 2.559	\$ 2.543

From Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006
Copyright – All rights reserved DATA WAREHOUSE: OLAP - 15 **Elena Baralis**
Politecnico di Torino

Database and data mining group, Politecnico di Torino
Slice and dice

The diagram illustrates 'slice and dice' operations. It shows a large rectangular area representing a data cube, with a vertical slice taken from it. This slice is then further divided into smaller rectangular sections, representing 'dice'.

- Selection of a data subset by means of selection predicates
 - slice: equality predicate selecting a “slice”
 - example: Year=2005
 - dice: predicate expression selecting a “dice”
 - example: Category='Food' and City='Torino'

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 16 **Elena Baralis**
Politecnico di Torino

Database and data mining group, Politecnico di Torino
DBG

Slice and dice

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

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 17 Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino
DBG

Slice and dice

Category	Year	Metric: Sales									
		Customer Region	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France
Electronics	1997	\$ 120	\$ 1774	\$ 204	\$ 128	\$ 2346	\$ 2554	\$ 2104	\$ 566	\$ 199	\$
	1998	\$ 1.184	\$ 4.529	\$ 1.990	\$ 7.220	\$ 651	\$ 9.498	\$ 476	\$ 2.662	\$ 460	\$ 7
Food	1997	\$ 759	\$ 692	\$ 725	\$ 265	\$ 560	\$ 469	\$ 807	\$ 156	\$ 615	\$ 1
	1998	\$ 949	\$ 820	\$ 920	\$ 292	\$ 712	\$ 1.020	\$ 1.020	\$ 1.020	\$ 1.020	\$ 1
Gifts	1997	\$ 2.542	\$ 1.335	\$ 1.894	\$ 1.413	\$ 3.320	\$ 2.182	\$ 1.804	\$ 649	\$ 376	\$ 1.0
	1998	\$ 1.988	\$ 2.735	\$ 2.800	\$ 2.659	\$ 1.813	\$ 2.944	\$ 2.729	\$ 1.156	\$ 717	\$ 8
Health & Beauty	1997	\$ 624	\$ 640	\$ 1.317	\$ 647	\$ 381	\$ 754	\$ 624	\$ 149	\$ 282	\$ 3
	1998	\$ 611	\$ 987	\$ 566	\$ 382	\$ 459	\$ 1.052	\$ 1.044	\$ 279	\$ 72	
Household	1997	\$ 9.254	\$ 4.112	\$ 9.410	\$ 4.444	\$ 3.254	\$ 3.974	\$ 2.454	\$ 2.045	\$ 2.075	\$ 1.9
	1998	\$ 10.191	\$ 9.191	\$ 9.491	\$ 4.467	\$ 4.466	\$ 3.949	\$ 2.464	\$ 2.056	\$ 2.056	\$ 2.7
Kid's Corner	1997	\$ 201	\$ 290	\$ 405	\$ 195	\$ 406	\$ 322	\$ 294	\$ 105	\$ 34	\$
	1998	\$ 247	\$ 422	\$ 441	\$ 380	\$ 221	\$ 592	\$ 290	\$ 198	\$ 39	\$
Travel	1997	\$ 624	\$ 505	\$ 564	\$ 366	\$ 300	\$ 978	\$ 416	\$ 40	\$ 20	
	1998	\$ 608	\$ 599	\$ 1.096	\$ 613	\$ 464	\$ 116	\$ 579	\$ 267	\$ 188	

Filter Criteria: Year = 1998		Metric: Sales								
Category	Customer Region	North-East	Mid-Atlantic	South-East	Central	South	North-West	England	France	Germany
Electronics	\$ 1.988	\$ 2.735	\$ 1.804	\$ 2.800	\$ 651	\$ 9.498	\$ 476	\$ 2.662	\$ 460	\$ 700
Food	\$ 1.973	\$ 2.785	\$ 2.800	\$ 2.485	\$ 1.813	\$ 2.844	\$ 1.778	\$ 1.238	\$ 717	\$ 1.000
Gifts	\$ 1.973	\$ 2.785	\$ 2.800	\$ 2.485	\$ 1.813	\$ 2.844	\$ 1.778	\$ 1.238	\$ 717	\$ 1.000
Health & Beauty	\$ 651	\$ 887	\$ 566	\$ 382	\$ 459	\$ 1.182	\$ 1.044	\$ 279	\$ 72	
Household	\$ 5.797	\$ 5.200	\$ 3.416	\$ 5.012	\$ 4.224	\$ 5.006	\$ 1.588	\$ 2.324	\$ 2.549	\$ 2.791
Kid's Corner	\$ 247	\$ 452	\$ 441	\$ 380	\$ 221	\$ 582	\$ 290	\$ 198	\$ 28	\$ 69
Travel	\$ 629	\$ 559	\$ 1.096	\$ 611	\$ 464	\$ 316	\$ 573	\$ 227	\$ 199	\$ 95

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

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 18 Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino
DB
M

Slice and dice

Metric	Customer	Dollar Sales	Alicon	Aileen	Albon	Alicameda	Aica	Allagash	Aica	Aicosia	Ametra	Amsterdam	Andersonville	Annas
Subcategory														
Audi								\$ 95						
Automobile									\$ 30					
Chocolate	\$ 45	\$ 42		\$ 50				\$ 20	\$ 22	\$ 44				
Cheesecake	\$ 35							\$ 25	\$ 30					
Class Tires								\$ 7	\$ 26					
Coffee									\$ 19					
Conduit									\$ 99					
Furniture										\$ 405				
Gadgets										\$ 199	\$ 79	\$ 79		
Games & Puzzles										\$ 17				
Gifts/Business											\$ 45			
Gifts	\$ 25													
Hearth											\$ 28	\$ 14		
Jewelry	\$ 75													
Kitchen														
Lanterns & Garden	\$ 75													
Lawn	\$ 15													
Meat & Cheese	\$ 15													
Miscellaneous	\$ 40													
Natural Remedies	\$ 200	\$ 1,320												
Parts	\$ 125													
Plants & Flowers	\$ 65	\$ 95	\$ 95											
Safety & Security														
Skin Care														
Sleeping														
Toys & Accessories														

↓

Metric	Customer	Dollar Sales	Alicon	Aileen	Albon	Alicameda	Aica	Allagash	Aica	Aicosia	Ametra	Amsterdam	Andersonville	Annas
Subcategory														
Audi								\$ 90		\$ 120	\$ 95			
comfort									\$ 118		\$ 1495			
Zeppetta														

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 19

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

*Elena Baralis
Politecnico di Torino*

Database and data mining group, Politecnico di Torino
DB
M

Pivot

- Reorganization of the multidimensional structure without varying the detail level
 - increases readability of the same information
 - multidimensional representation is always based on a “grid” (hierarchical spreadsheet)
 - two dimensions are the main grid axes
 - position of dimensions in the grid are changed

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 20

*Elena Baralis
Politecnico di Torino*

Database and data mining group, Politecnico di Torino
DBG

Pivot

From Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006
Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 21

*Elena Baralis
Politecnico di Torino*

Database and data mining group, Politecnico di Torino
DBG

Pivot

Category	Metrics	Dollar Sales
Electronics	Year	
Electronics	1997	\$ 10.615
Electronics	1998	\$ 29.299
Food	1997	\$ 5.300
Food	1998	\$ 5.626
Gifts	1997	\$ 16.215
Gifts	1998	\$ 20.017
Health & Beauty	1997	\$ 4.042
Health & Beauty	1998	\$ 5.665
Household	1997	\$ 38.383
Household	1998	\$ 50.391
Kids' corner	1997	\$ 2.559
Kids' corner	1998	\$ 2.943
Travel	1997	\$ 4.497
Travel	1998	\$ 4.792

↓

Metrics	Year	Dollar Sales
Category		
Electronics	1997	\$ 10.615
Electronics	1998	\$ 29.299
Food	1997	\$ 5.300
Food	1998	\$ 5.626
Gifts	1997	\$ 16.215
Gifts	1998	\$ 20.017
Health & Beauty	1997	\$ 4.042
Health & Beauty	1998	\$ 5.665
Household	1997	\$ 38.383
Household	1998	\$ 50.391
Kids' corner	1997	\$ 2.559
Kids' corner	1998	\$ 2.943
Travel	1997	\$ 4.497
Travel	1998	\$ 4.792

From Golfarelli, Rizzi, "Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006
Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 22

*Elena Baralis
Politecnico di Torino*

Database and data mining group, Politecnico di Torino


Pivot

Category	Year	Dollar Sales											
		Metro	Customer	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany
Region													
Electronics	1997			\$ 120	\$ 1.774	\$ 204	\$ 130	\$ 2.346	\$ 2.254	\$ 2.104	\$ 566	\$ 159	\$
	1998			\$ 1.104	\$ 4.529	\$ 1.090	\$ 7.220	\$ 8.651	\$ 9.498	\$ 4.760	\$ 2.662	\$ 462	\$ 7
Food	1997			\$ 759	\$ 652	\$ 725	\$ 265	\$ 561	\$ 469	\$ 607	\$ 156	\$ 615	\$ 1
	1998			\$ 538	\$ 925	\$ 259	\$ 677	\$ 213	\$ 1.920	\$ 261	\$ 165	\$ 175	\$ 1
Gifts	1997			\$ 3.532	\$ 1.795	\$ 1.894	\$ 1.477	\$ 2.525	\$ 2.192	\$ 1.094	\$ 908	\$ 376	\$ 1.8
	1998			\$ 3.089	\$ 2.099	\$ 2.689	\$ 1.961	\$ 1.715	\$ 1.841	\$ 1.040	\$ 916	\$ 277	\$ 8
Health & beauty	1997			\$ 624	\$ 640	\$ 1.017	\$ 647	\$ 288	\$ 734	\$ 524	\$ 143	\$ 124	\$ 3
	1998			\$ 611	\$ 887	\$ 566	\$ 282	\$ 459	\$ 1.062	\$ 564	\$ 124	\$ 143	\$ 3
Household	1997			\$ 9.354	\$ 4.115	\$ 9.410	\$ 4.444	\$ 3.856	\$ 3.974	\$ 2.654	\$ 3.945	\$ 2.675	\$ 1.9
	1998			\$ 5.767	\$ 5.220	\$ 5.410	\$ 6.012	\$ 4.324	\$ 5.008	\$ 7.500	\$ 2.126	\$ 3.649	\$ 2.7
Kids' corner	1997			\$ 201	\$ 390	\$ 405	\$ 196	\$ 409	\$ 222	\$ 204	\$ 105	\$ 34	\$
	1998			\$ 247	\$ 422	\$ 441	\$ 365	\$ 221	\$ 592	\$ 290	\$ 106	\$ 370	\$
Travel	1997			\$ 624	\$ 505	\$ 564	\$ 266	\$ 300	\$ 971	\$ 465	\$ 46	\$ 20	
	1998			\$ 608	\$ 799	\$ 1.094	\$ 613	\$ 464	\$ 918	\$ 579	\$ 207	\$ 196	

↓

Category	Year	Dollar Sales													
		Metro	Customer	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany		
Region															
Electronics	1997			\$ 130	\$ 1.104	\$ 1.774	\$ 4.529	\$ 304	\$ 1.092	\$ 130	\$ 7.220	\$ 2.346	\$ 651	\$ 2.554	\$ 9.480
	1998			\$ 759	\$ 538	\$ 652	\$ 725	\$ 265	\$ 677	\$ 200	\$ 223	\$ 469	\$ 156	\$ 615	\$ 1.502
Food	1997			\$ 538	\$ 759	\$ 652	\$ 725	\$ 265	\$ 677	\$ 200	\$ 223	\$ 469	\$ 156	\$ 615	\$ 1.502
	1998			\$ 258	\$ 624	\$ 640	\$ 1.017	\$ 647	\$ 444	\$ 1.094	\$ 1.040	\$ 1.040	\$ 376	\$ 124	\$ 124
Gifts	1997			\$ 3.532	\$ 624	\$ 640	\$ 1.017	\$ 647	\$ 444	\$ 1.094	\$ 1.040	\$ 376	\$ 124	\$ 124	\$ 1.8
	1998			\$ 3.089	\$ 611	\$ 887	\$ 566	\$ 288	\$ 459	\$ 1.062	\$ 564	\$ 124	\$ 143	\$ 143	\$ 1.8
Health & beauty	1997			\$ 624	\$ 611	\$ 887	\$ 566	\$ 288	\$ 459	\$ 1.062	\$ 564	\$ 124	\$ 143	\$ 143	\$ 1.8
	1998			\$ 611	\$ 608	\$ 799	\$ 1.094	\$ 613	\$ 464	\$ 918	\$ 579	\$ 207	\$ 196		
Household	1997			\$ 9.354	\$ 5.767	\$ 4.115	\$ 5.220	\$ 5.410	\$ 6.012	\$ 2.654	\$ 3.945	\$ 2.675	\$ 1.9		
	1998			\$ 5.767	\$ 608	\$ 799	\$ 1.094	\$ 613	\$ 464	\$ 918	\$ 579	\$ 207	\$ 196		
Kids' corner	1997			\$ 201	\$ 247	\$ 390	\$ 422	\$ 441	\$ 365	\$ 222	\$ 204	\$ 105	\$ 34		
	1998			\$ 247	\$ 201	\$ 390	\$ 422	\$ 441	\$ 365	\$ 222	\$ 204	\$ 105	\$ 370		
Travel	1997			\$ 624	\$ 624	\$ 505	\$ 564	\$ 1.092	\$ 300	\$ 971	\$ 465	\$ 46	\$ 20		
	1998			\$ 608	\$ 624	\$ 505	\$ 564	\$ 1.092	\$ 300	\$ 971	\$ 465	\$ 46	\$ 20		

From Golfarelli, Rizzi,"Data warehouse, teoria e pratica della progettazione", McGraw Hill 2006
Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 23

Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino


Extensions of the SQL language

- Interface tools require
 - new aggregate functions
 - aggregate functions exploited for economic analysis (moving average, median, ...)
 - position in the sort order (i.e., rank)
 - functions for report generation
 - partial and cumulative totals
- New OLAP functions in the ANSI standard
 - implemented starting from DB2 UDB 7.1, Oracle 8i v2

Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 24

Elena Baralis
Politecnico di Torino

Extensions of the SQL language

- Interface tools require
 - operators for the computation of different group bys at the same time
- The SQL-99 (SQL3) standard has extended the SQL group by clause

Example data base

Sales(City,Month,Amount)

City	Month	Amount
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

SQL OLAP functions

- New class of aggregate functions (OLAP functions) characterized by
 - computation window, inside which the computation of aggregate functions is performed
 - cumulative totals and moving average can be computed
 - new aggregate functions to compute the rank in a given sort order

Computation window

- New **window** clause, characterized by
 - *partitioning*: Rows are grouped without collapsing them (different from **group by**)
 - no partitioning: a single group is defined
 - *row ordering*, separately in each partition (similar to **order by**)
 - *aggregation window*: For each row in the partition, it defines the row group on which the aggregate function is computed

Example

- Show, for each city and month
 - sale amount
 - average on the current month and the two previous months, separately for each city

Example

- Partitioning on city
 - average computation is reset when the city changes
- Ordering by month, to compute the moving average on the current month and the two preceding months
 - without ordering the computation is meaningless
- Aggregation window size: the current row and the two preceding rows

Example

```

SELECT City, Month, Amount,
       AVG(Amount) OVER Wavg AS MovingAvg
  FROM Sales
 WINDOW Wavg AS (PARTITION BY City
                  ORDER BY Month
                  ROWS 2 PRECEDING)
    
```

Example

```

SELECT City, Month, Amount,
       AVG(Amount) OVER (PARTITION BY City
                          ORDER BY Month
                          ROWS 2 PRECEDING)
              AS MovingAvg
  FROM Sales
    
```

Database and data mining group, Politecnico di Torino
DBM

Result

City	Month	Amount	MovingAvg
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

Partition 1

Partition 2

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 33 Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino
DBM

Observations

- Sort order is required, because the computation of the moving average considers rows in an ordered fashion
 - the window sort order does not enforce a predefined output sort order
- When the window is not complete, the computation takes place on the available rows
 - it is possible to require a **NULL** result for each incomplete window
- Several different computation windows may be specified

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 34 Elena Baralis
Politecnico di Torino

Aggregation window

- The moving window on which the aggregate function is computed may be defined
 - at the *physical level*: It builds the group by counting rows
 - example: the current row and the two preceding rows
 - at the *logical level*: It builds the group by defining an interval on the sort key
 - example: the current month and the two preceding months

Physical interval definition

- Between a lower bound and the current row
`ROWS 2 PRECEDING`
- Between lower and upper bounds
`ROWS BETWEEN 1 PRECEDING AND 1 FOLLOWING`
`ROWS BETWEEN 3 PRECEDING AND 1 PRECEDING`
- Between the beginning (or the end) of a partition and the current row
`ROWS UNBOUNDED PRECEDING (o FOLLOWING)`

Physical interval

- Appropriate for sequence data with no gaps
 - example: no month is missing in the sequence
 - more than a sort key can be specified
 - computation ignores breaks due to change in any sort key value
 - example: order by month and year
 - no mathematical expressions are needed to compute the window

Logical interval definition

- The **range** clause is used, with the same syntax as the physical interval
- A distance on the sort key between the interval bounds and the current value should be defined
- Example

RANGE 2 MONTH PRECEDING

Logical interval

- Appropriate for “sparse” data, with gaps in the sequence
 - example: a month is missing in the sequence
 - only a single sort key can be specified
 - the sort key can only be alphanumeric or date type (arithmetic expressions are allowed)

Applications

- Moving aggregate computations
 - computations on a window which moves over data
 - examples: moving average, moving sum
- Cumulative total computations
 - the (cumulative) total is incremented by adding an instance at a time
- Comparison between detailed data and aggregated data

Computation of a cumulative total

- Show, for each city and month
 - sale amount
 - cumulative sale amount for increasing months, separately for each city

Computation of a cumulative total

- Partition by city
 - the cumulative total is reset when the city changes
- Order by (ascending) month to compute the sum for increasing months
 - without sorting, the computation would be meaningless
- Size of the aggregation window
 - from the starting row of the partition to the current row

Database and data mining group, Politecnico di Torino
DBM

Computation of a cumulative total

```

SELECT City, Month, Amount,
       SUM(Amount) OVER (PARTITION BY City
                          ORDER BY Month
                          ROWS UNBOUNDED PRECEDING)
          AS CumeTot
  FROM Sales
    
```

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 43 Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino
DBM

Computation of a cumulative total

City	Month	Amount	CumeTot
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

Partition 1 Partition 2

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 44 Elena Baralis
Politecnico di Torino

Comparison between detailed data and total data

- Show, for each city and month
 - sale amount
 - total sale amount on the whole time period for the current city

Comparison between detailed data and total data

- Partition by city
 - the total amount is reset when the city changes
- Sorting is not needed
 - the total amount is computed independently of the sort order of tuples
- The aggregation window is not needed
 - it is the whole partition

Database and data mining group, Politecnico di Torino
DBM

Comparison between detailed data and total data

```

SELECT City, Month, Amount,
       SUM(Amount) OVER (PARTITION BY City)
       AS TotalAmount
FROM Sales
    
```

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 47 Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino
DBM

Comparison between detailed data and total data

City	Month	Amount	TotalAmount
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

Elena Baralis
Politecnico di Torino

Partition 1
Partition 2

Copyright – All rights reserved DATA WAREHOUSE: OLAP - 48 Elena Baralis
Politecnico di Torino

Comparison between detailed data and total data

- Show, for each city and month
 - sale amount
 - ratio between current row amount and grand total
 - ratio between current row amount and total amount by city
 - ratio between current row amount and total amount by month

Comparison between detailed data and total data

- Three different computation windows
 - grand total: no partitioning
 - total by city: partition by city
 - total by month: partition by month
- No sort is needed in any window
 - totals are independent of the sort order of tuples
- The aggregation window is always the whole partition

Database and data mining group, Politecnico di Torino



Comparison between detailed data and total data

```

SELECT City, Month, Amount
    Amount/SUM(Amount) OVER () AS TotalFract
    Amount/SUM(Amount) OVER (PARTITION BY City) AS CityFract
    Amount/SUM(Amount) OVER (PARTITION BY Month) AS MonthFract
FROM Sales
    
```

Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 51

Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino



Comparison between detailed data and total data

City	Month	Amount	TotalFract	CityFract	MonthFrct
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

Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 52

Elena Baralis
Politecnico di Torino

Group by and window

- Windows can be used together with grouping performed by **group by**
- The “temporary table” generated by the execution of the **group by** clause (possibly with aggregate function computation) becomes the operand to which the computations in the **window** clause are applied

Example

- Assume that the **Sales** table contains information on sales with daily granularity
- Show, for each city and month
 - sale amount
 - average sale with respect to the current month and the two preceding months, separately for each city

Example

- Grouping by month is needed to compute the total amount by month before computing the moving average
 - the group by clause is used for computing the monthly total
- The temporary table generated by the group by computation is the operand on which the computation window is defined

Example

```

SELECT City, Month, SUM(Amount) AS TotMonth,
       AVG(SUM(Amount)) OVER (PARTITION BY City
                               ORDER BY Month
                               ROWS 2 PRECEDING)
           AS MovingAvg
  FROM Sales
 GROUP BY City, Month
    
```

Ranking functions

- Functions computing the rank of a value inside a partition
 - **rank()** function: computes the rank by leaving an empty slot after a tie
 - example: after 2 first, the next rank is third
 - **denserank()** function: computes the rank by leaving an empty slot after a tie
 - example: after 2 first, the next rank is second

Example

- Show, for each city in december
 - sale amount
 - rank on amount

Example

- Partitioning is not needed
 - a single partition including all cities
- Order by amount to perform ranking
 - without sorting, the computation would be meaningless
- The aggregation window is the whole partition

Example

```
SELECT City, Amount,
       RANK() OVER (ORDER BY Amount DESC)
       AS Ranking
FROM Sales
WHERE Month = 12
```

Result

City	Amount	Ranking
Torino	150	1
Milano	140	2

Sorting the result

- A sorted result is obtained by means of the **order by** clause
 - may be different from the sort order in the computation window
- Example: sort the result in the former example by increasing city

Example

```
SELECT City, Amount,
       RANK() OVER (ORDER BY Amount DESC)
AS Ranking
FROM Sales
WHERE Month = 12
ORDER BY City
```

City	Amount	Ranking
Milano	140	2
Torino	150	1

group by clause extensions

- Multidimensional spreadsheets compute several partial totals “in one shot”
 - total sale amount by month and city
 - total sale amount by month
 - total sale amount by city
- For the sake of efficiency avoid
 - multiple data reads
 - redundant data sorts

group by clause extensions

- SQL-99 standard extended the syntax of the **group by** clause
 - **rollup** computes aggregations on all groups obtained by removing one by one the columns in the grouping clause
 - **cube** computes aggregations on all combinations of the columns in the grouping clause
 - **grouping sets** computes aggregations on the group list in the grouping clause (grouping sets different from the previous clauses may be specified)
 - () for grand totals (no grouping)

Rollup: example

- Consider the following tables


```
Time(Tkey, Day, Month, Year, ...)  
Shop(Skey, City, Region, ...)  
Product(Pkey, PName, Brand, ...)  
Sales(Skey, Tkey, Pkey, Amount)
```
- Compute total sales in the year 2000 for the following attribute combinations
 - product, month, city
 - month, city
 - city

Rollup: example

```

SELECT City, Month, Pkey,
       SUM(Amount) AS TotSales
  FROM Time T, Shop S, Sales V
 WHERE T.Tkey = V.Tkey
   AND S.Skey = V.Skey
   AND Year = 2000
 GROUP BY ROLLUP (City,Month,Pkey)
    
```

- The column sort order in **rollup** determines which aggregates are computed

Rollup: result

City	Month	Pkey	TotSales
Milano	7	145	110
Milano	7	150	10
Milano
Milano	7	NULL	8500
Milano	8
Milano	NULL	NULL	150000
Torino	150
Torino	...	NULL	2500
Torino	NULL	NULL	135000
...
NULL	NULL	NULL	25005000

- “Superaggregates” are represented by **NULL**

Cube: example

- Compute total sales in the year 2000 for *all* combinations of the following attributes
 - product, month, city
- The following aggregations should be computed
 - product, month, city
 - product, month
 - month, city
 - product, city
 - product
 - month
 - city
 - no grouping

Cube: example

```

SELECT City, Month, Pkey,
       SUM(Amount) AS TotSales
  FROM Time T, Shop S, Sales V
 WHERE T.Tkey = V.Tkey
   AND S.Skey = V.Skey
   AND Year = 2000
 GROUP BY CUBE (City,Month,Pkey)
    
```

- The sort order of columns in **cube** is irrelevant

Cube computation

- Consider distributive and algebraic properties of aggregate functions
 - *distributive* aggregate functions (**min**, **max**, **sum**, **count**) may be computed from aggregations on a larger set of attributes (i.e., with larger granularity)
 - Example: from total sales by product and month, total sales by month may be computed
 - *algebraic* aggregate functions (**avg**, ...) may be computed from aggregations on a larger set of attributes (i.e., with larger granularity), if appropriate support aggregations are stored
 - Example: average requires
 - the average value in the group
 - the cardinality of the group

Cube computation

- To increase the efficiency of cube computation, the distributive/algebraic properties of the aggregate functions are exploited
 - previously computed **group by** are exploited
 - **rollup** requires a single sort operation
 - the cube is a combination of several **rollup** operations (in the appropriate order)
 - previously executed sort operations are exploited (also partially)
 - it is possible to exploit sort on (A,B) to sort by (A,C)

Database and data mining group, Politecnico di Torino



Grouping Set: example

- Compute total sales in the year 2000 for the following groups
 - month
 - month, city, product
- A roll up would perform the computation of unnecessary groupings and aggregations

Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 73

Elena Baralis
Politecnico di Torino

Database and data mining group, Politecnico di Torino



Grouping Set: example

```
SELECT City, Month, Pkey,
       SUM(Amount) AS TotSales
  FROM Time T, Shop S, Sales S
 WHERE T.Tkey = S.Tkey
   AND S.Skey = S.Skey
   AND Year = 2000
 GROUP BY GROUPING SETS
        ((Month), (City, Month, Pkey))
```

Copyright – All rights reserved

DATA WAREHOUSE: OLAP - 74

Elena Baralis
Politecnico di Torino