UNIVERSITÀ
DEGLI STUDI DI TRIESTE

Data WareHouse - Introduction

Prof. A. Peron

Slides from M. Golfarelli, S. Rizzi,
Datawarehouse Design, Modern Principles and methodologies, McGrawHill.

(Slightly modified by Dario Della Monica)

Data Warehouse

- Data warehouse is a repository for historical integrated and consistent data.
- it is equipped with the tools that offer company management the opportunity to extract reliable information to be used as a support for the decision making process
- data warehousing involves processes that
  - extract the relevant data from an enterprise information system
  - transform the data, integrate it, remove any flows and inconsistencies
  - store it into a data warehouse
  - provide end users with access to the data; they can carry out complex data analysis and prediction queries

Data Warehouse

- Information assets are immensely valuable to any enterprise
- This assets must be properly stored and readily accessible when they are needed.
- Data warehousing is a phenomenon that grew from
  - the huge amount of electronic data stored in recent years
  - the urgent need to use the data to accomplish goals that go beyond the routine task linked to daily processing

Some Fields of application:

- **Trade:** sales, shipment and inventory analysis, customer care and public relations
- **Craftsmanship:** production cost control, suppliers, and order support
- **transport industry:** vehicle management
- **telecommunications services:** call flow analysis and customer profile analysis
- **healthcare service:** patient admission and discharge, analysis and bookkeeping in department accounts

Decision support systems

- Traditional approach to data management:
- **OLTP** (On Line Transaction Processing): Storing operational data managed by transactions.

- Data management for Decision support
- **OLAP** (On Line Analytical Processing): transform operational data into decision-making support information

Requirementes for a DW process

- **Data warehousing process**: set of tasks that allow us to turn operational data into decision making support information
- Requirements
- **Accessibility:** to users are not very familiar with the IT and data structures;
- **Integration all the data on the basis of a standard enterprise model;**
- **Query flexibility** to maximize the advantages obtained from the existing information;
- **Information conciseness** allowing for target oriented and effective analyses;
- **Multidimensional representation giving users an intuitive and manageable view of information;**
- **Correctness and completeness** of integrated data.

DataWarehousing (def. 2)

▶**Data warehouse**. A Data Warehouse (DW) is a collection of data that supports decision making processes. It provides the following features:

- it is subject oriented;
- it is integrated and consistent;
- it shows its evolution over time and it is not volatile.

- It is based on operational (possibly) heterogeneous data sources;
- Does not require that new information be added; rather existing information needs rearranging;
- Are regularly updated from operational data and keep on growing (big collections of data)
- Data is never deleted from data warehouses and updates are normally carried out when data warehouses are offline
- …

DataWarehousing (def. 2)

- …
- Data warehouse can be essentially viewed as read only databases
- There is no need for advanced transaction management techniques usually required by operational applications
- Table normalization can be given up to partially denormalized tables and improve performance
- Queries feature dynamic and multidimensional analyses involving a huge amount of records

DW vs operational DataBase

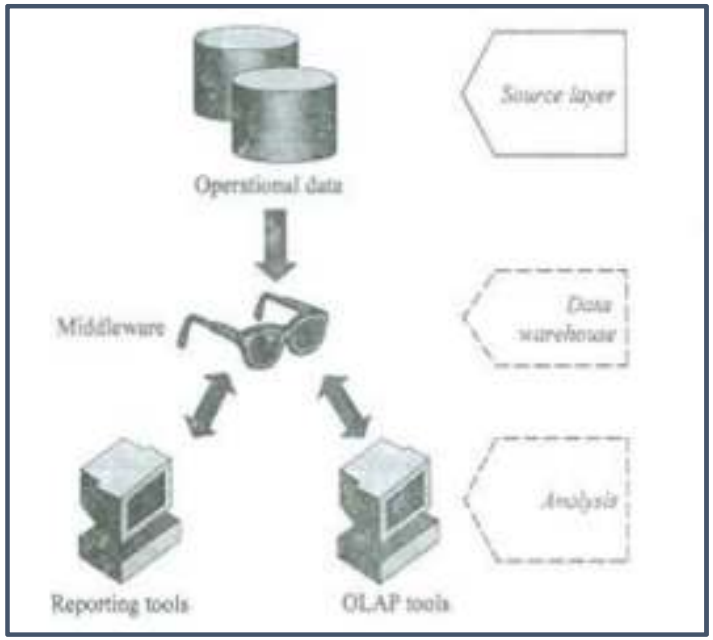
Feature	Operational Databases	Data Warehouses
Users	Thousands	Hundreds
Workload	Preset transactions	Specific analysis queries
Access	To hundreds of records, write and read mode	To millions of records, mainly read-only mode
Goal	Depends on applications	Decision-making support
Data	Detailed, both numeric and alphanumeric	Summed up, mainly numeric
Data integration	Application-based	Subject-based
Quality	In terms of integrity	In terms of consistency
Time coverage	Current data only	Current and historical data
Updates	Continuous	Periodical
Model	Normalized	Denormalized, multidimensional
Optimization	For OLTP access to a database part	For OLAP access to most of the database

TABLE 1-2 Differences Between Operational Databases and Data Warehouses (Kelly, 1997)

DW Architectures

- **essential properties of DW architectures.**
- **Separation:** analytical and transactional processing should be kept apart as much as possible
- **Scalability:** hardware and software architectures should be easy to upgrade as the data volume which has to be managed progressively increases
- **Extensibilty:** the architecture should be able to host new application (tools for data visualization) without redesigning the whole system
- **Security:** monitoring accesses is essential because of the strategic data stored in data warehouses
- **Administerability:** data warehouse management should not be difficult

Single-layer Architetture



Single-layer architecture

- Virtual DW implemented as a **multidimensional view of the operational** DB.
- **Separation requirement is not fulfilled**
- The analytical tasks interfere with the transactional tasks.
- It can be adopted only if the analytical task is restricted.

Two layer Architecture

Two-layer architecture

- The separation requirement is guaranteed.
- **Source layer:**
  - heterogeneous sources of data (relational databases, legacy databases, information systems outside the corporate wall)
- **Data staging:**
- **The data storage to sources should be**
  - **extracted**
  - **cleansed to remove inconsistencies and fill gaps and**
  - **integrated to merge heterogeneous sources into one common schema.**

ETL stage (Extraction, Transformation and Loading).

Two-layer architecture (2)

- **Data warehouse layer:**
- Information is stored to one logically centralized single repository
- the data warehouse can be directly accessed
- it can also be used as a source for creating DATA MARTS which partially replicate data warehouse contents and are designed for specific enterprise department.
- **Metadata** keep information on:
  - data sources,
  - access procedures,
  - cleansing procedures
  - loading,
  - Data mart schemata .

Two-Layers Architecture (3)

- **Analysis level:**
- Integrated data is efficiently and effectively accessed to
  - issue reports
  - dynamically analyzing information and
  - simulate hypothetical business scenarios
- **From the technological viewpoint it should feature**
  - **aggregated data navigators,**
  - **complex query optimizers and**
  - **user friendly GUIs**

Datamart

- It's a subset or an aggregation of data stored to a primary data warehouse; it includes a set of information pieces relevant to a specific business area, corporate department or category of users.
- Dependent Data marts:
- **populated from a primary data warehouse (they replicate a part of the data warehouse)**
- Useful (non strictly necessary) since :
  - **They mark out the information required by a specific group of users**
  - **They can deliver better performances because they are smaller**
- Independent Data marts:
- They are directly populated from operational sources
  - a primary data warehouse is lacking …
  - … or it can be built merging all Data marts (bottom-up design)

Two layer architecture with independent data marts

Benefits of a two-layer architecture

- That information is always available even when the access to sources is denied
- The analysis queries do not affect the management of transactions
- Are logically structured according to the multidimensional model (operational sources are generally based on relational or semi-structured models)
- Allow to manage the proper level of time and granularity:
  - **OLTP**: deal with data at the greatest detail level.
  - **OLAP manage historical and summarized data**.
- Specific design solution aimed at optimizing performance of analysis and report applications.

Three-layer Architetture

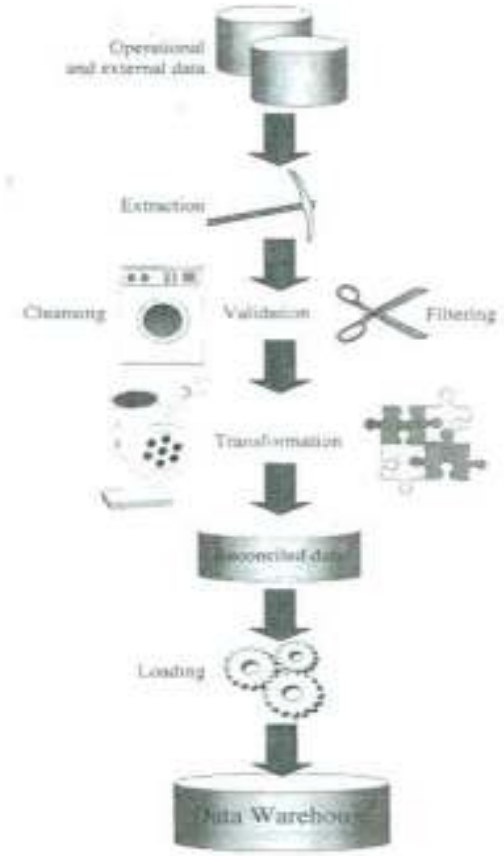
Three-layers architecture

- **The level of reconciled data is added**
- **This layer materializes operational data obtained after integrating and cleansing sourced data.**
- The DW is populated from the reconciled data level instead being populated from the operational data.
- Benefits:
- It creates a common reference data model for a whole enterprise
- It sharply separates the problems of source data **extraction** and integration (**transformation**) from those of the data warehouse population (**loading**)

Data Staging and ETL

- **The data staging layer hosts the ETL processes that extract, integrate and clean data from operational sources**
- **Extraction, Transformation, Loading**
- The most complex and technically challenging among all the data warehouse process phases.
- ETL takes place:
- once when the data warehouse is populated for the first time
- Every time the data warehouse is regularly updated

Data Staging and ETL



Extraction

- **Relevant data are extracted from the operational sources**
- **Static extraction**: when a data warehouse needs populating for the first time
- **Incremental extraction**: used to populate the data regularly. It reflects the changes applied to source data since the last extraction
  - **Incremental** extraction can be based on:
  - **"journaling" (log)** managed by the operational DBMS;
  - **Time-stamping** when the operational data are provided with a temporal mark when inserted or updated.
  - **Source-driven:** operational applications asynchronously notify the changes in the operational sources (e.g. triggers combined with transactions).

Cleansing

- The cleansing phase is crucial in a data warehouse system because it is supposed to improve the data quality normally quite poor in external sources and not uniform across different internal operation sources.

- Most frequent mistakes and inconsistency of data:

- **duplicate data**;
  - e.g. A patient is recorded many times in an hospital information system.

- **Inconsistent values that are logically associated;**
  - e.g. addresses and ZIP codes

- **Missing data**;

- **Unexpected use of fields**
  - e.g. a Social Security number field could be used improperly to store office phone numbers.

- **Impossible or wrong values** (e.g. '31/09/2015');

Cleansing (2)

- **Inconsistent values for a single line entity because different practices were used**
  - e.g. an international country abbreviation (I) or a full country name (Italy);
- **Inconsistent values for one individual entity because of a typing mistakes**
  - e.g. Hamet Road instead of Hamlet Road

The main data cleansing features found in ETL tools are **rectification** and almost **homogenization**

they use of specific **dictionaries** to rectify typing mistakes and thesauri to recognize synonyms

rule based cleansing to enforce domain specific rules and define appropriate association between values

Transformation

- It converts the data from its operational source format into a specific data warehouse format.
- Establishing a mapping between the source of data layer and the data warehouse layer is generally made difficult by the presence of many different heterogeneous sources.
- **Examples of Transformations:**
  - Extraction of a structured information from a text string (e.g. the fields of an address encoded by a string)
  - Normalization in the usage of frequently used data types (e.g. a date may be encoded as a string or as a triple of integers)
- **Transformation processes:**
- **Conversion and normalization:** they operate on both storage formats and units of measure to make data uniform
- **Matching**: it associates equivalent fields in different sources
- **Selection**: it reduces the number of source fields and the records.

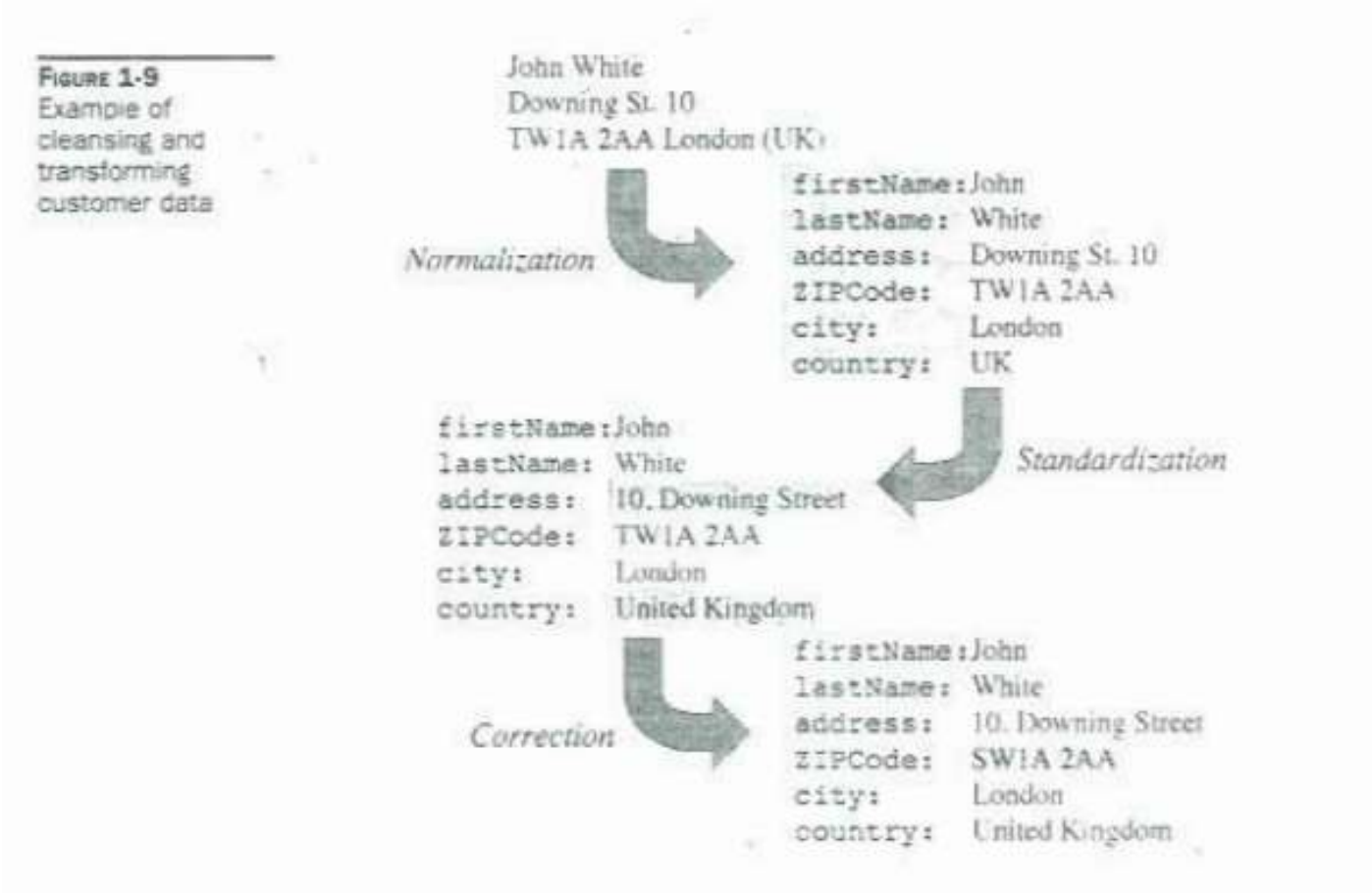


Figure 1-9
Example of cleansing and transforming customer data

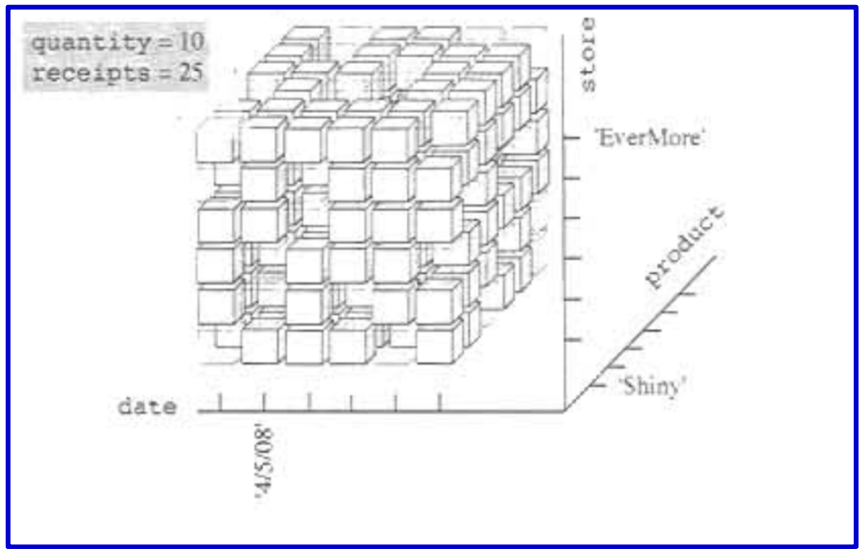
Multidimensional model

- Factors affecting the decision making processes are enterprise specific factors
  - (e.g. sales, shipments, hospital admissions or surgeries etc.)
- Instances of a fact correspond to events that occurred
- Each factor is described by the values of a set of relevant measures that provide a quantitative description of events
  - (e.g. Amount shipped, hospital admission costs, surgery time are measures)
- Events have a huge cardinality (milions/bilions)
- Events are placed into a n-dimensional space to help quickly select and sort out.
- The n-dimensional space axes are called analysis dimensions and give the final different perspectives to single out or aggregate events
  - (e.g. Sales in a store chain can be represented in a three-dimensional space whose dimensions are products, stores and dates)

Multidimensional model

- The concept of dimension prides the metaphor of cubes to represent multidimensional data
- Events are associated with the cube cells and the cube edges stand for analysis dimensions.
- The relational schema for the cube would be

SALES(store,product,date,quantity,receipts)



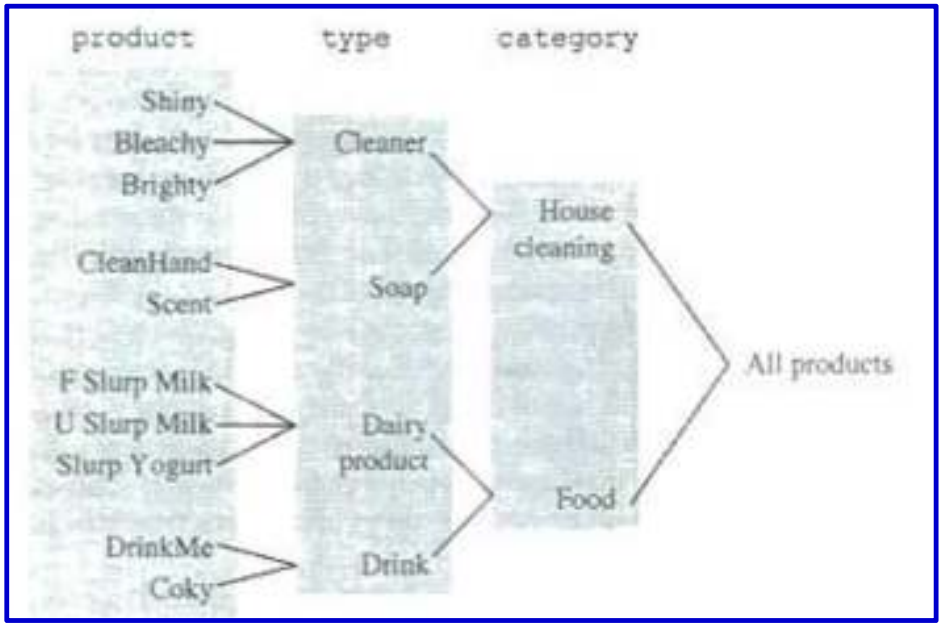
Multidimensional model

- Each dimension is associated with a hierarchy of aggregation levels often called roll up hierarchy.
- Roll up hierarchies group aggregation levels value in different ways
- Hierarchies consist of levels called dimensional attributes

  (E.g. product → type → category

  store → City → State)

- In summary each dimension can be analyzed at a different detail levels specified by hierarchical structured attributes

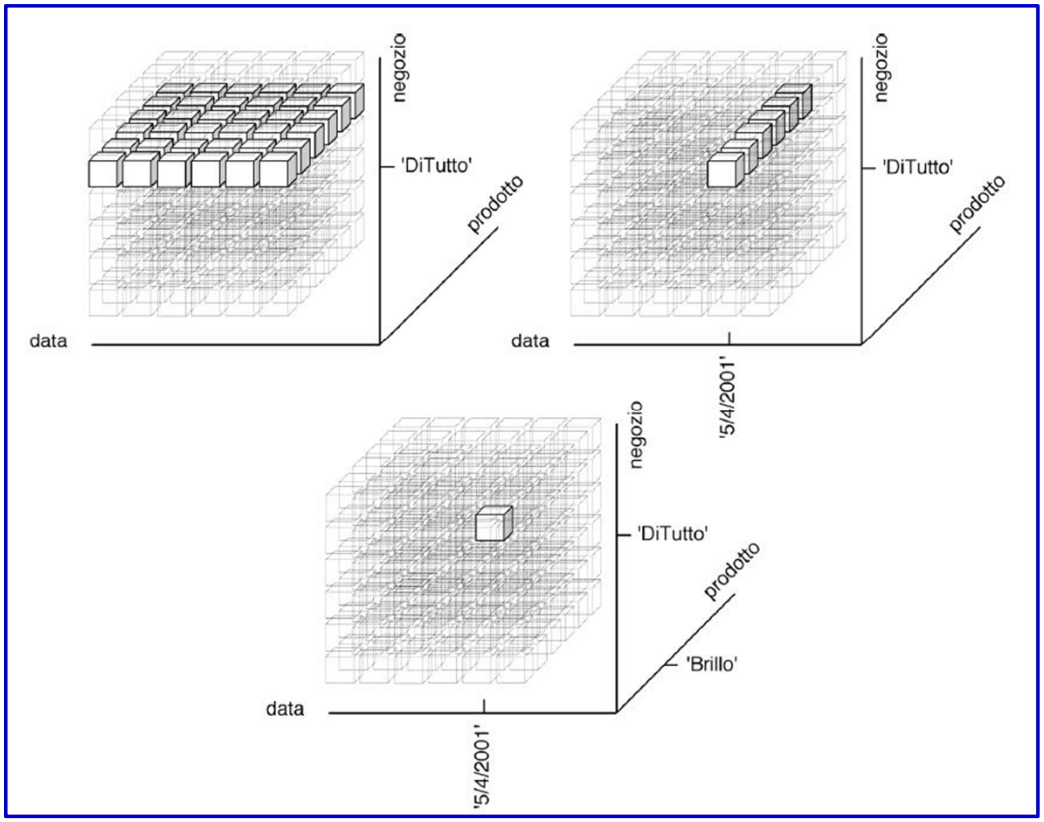


Multidimensional model

- Information in a multi-dimensional cube is very difficult for users to manage because of its quantity.
- E.g. a cube for 50 stores selling **1000 products**, and **3 years of transactions** (≈ 1000 days) has possibly **50×1000×1000=5×10^7 events**.
- Assuming that each store can sell only 10% of all the available items per day the number of events totals **5×10^6**
- The way to reduce the quantity of data and obtain useful informations are
- **restriction:** Separating a part of the data from a cube to mark out and analysis field (selection/projection);
- **aggregation:** reduces the granularity of the cube by exploiting the hierarchy on dimensions.

Restriction

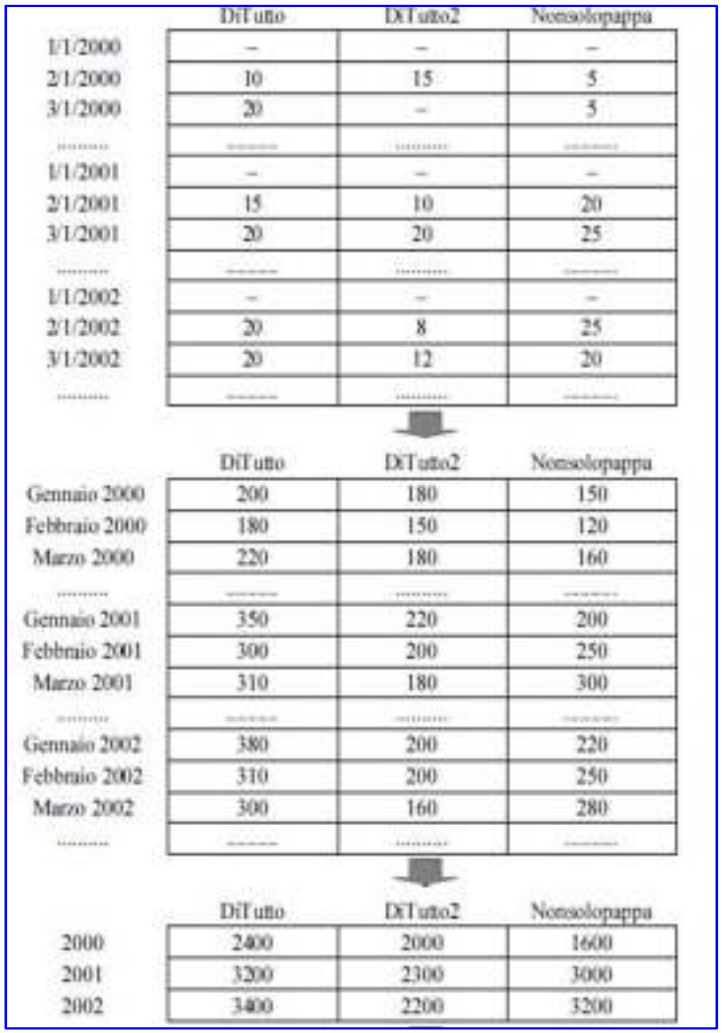
- **Slicing**. Decreases the cube dimensionality by settung one or more dimensions to a specific value.
- Projection. A choice to consider only a subset of the mesures of events.
- Dicing. A generalization of Slicing. It fixes a subset of values of a dimensional attribute



Aggregation on a dimension hierarchy

Aggregation:

- cells are aggregated with respect to one or more dimensional attributes
- Every aggregate event sums up the measures associated with events it aggregates.
- E.g. in figure events are aggregated along the temporal dimension: first by month and then by year.



	DiTutto	DiTutto2	Nonsolopappa
1/1/2000	–	–	–
2/1/2000	10	15	5
3/1/2000	20	–	5
.....	.....	.....	.....
1/1/2001	–	–	–
2/1/2001	15	10	20
3/1/2001	20	20	25
.....	.....	.....	.....
1/1/2002	–	–	–
2/1/2002	20	8	25
3/1/2002	20	12	20
.....	.....	.....	.....

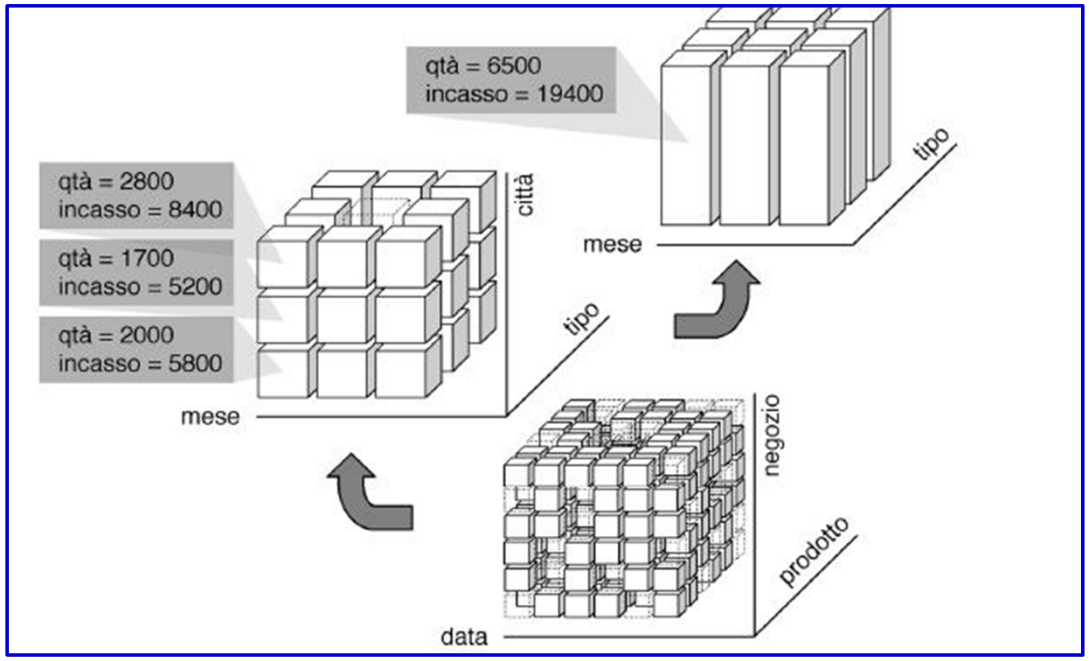
	DiTutto	DiTutto2	Nonsolopappa
Gennaio 2000	200	180	150
Febbraio 2000	180	150	120
Marzo 2000	220	180	160
.....	.....	.....	.....
Gennaio 2001	350	220	200
Febbraio 2001	300	200	250
Marzo 2001	310	180	300
.....	.....	.....	.....
Gennaio 2002	380	200	220
Febbraio 2002	310	200	250
Marzo 2002	300	160	280
.....	.....	.....	.....

	DiTutto	DiTutto2	Nonsolopappa
2000	2400	2000	1600
2001	3200	2300	3000
2002	3400	2200	3200

Example: two levels of aggregation

Two aggregation steps

1. By month
2. By type



Analysis: Querying a data warehouse

- How end users query a data warehouse:
- **Reports:** approach oriented to those users who need to have regular access to the information in almost static way.
- The layout of the report is predetermined.
- A fixed set of queries are predetermined to create reports with the desired layout and freeze all those in an application.
- A layout can look like a table or a chart (diagrams, histograms pies and so on)
- **OLAP**
- OLAP is the main way to exploit information in a data warehouse
- it gives the opportunity to analyze and explore data interactively on the basis of the multidimensional model
- Users are able to start a complex analysis session actively where each step is the result of the outcome of preceding steps

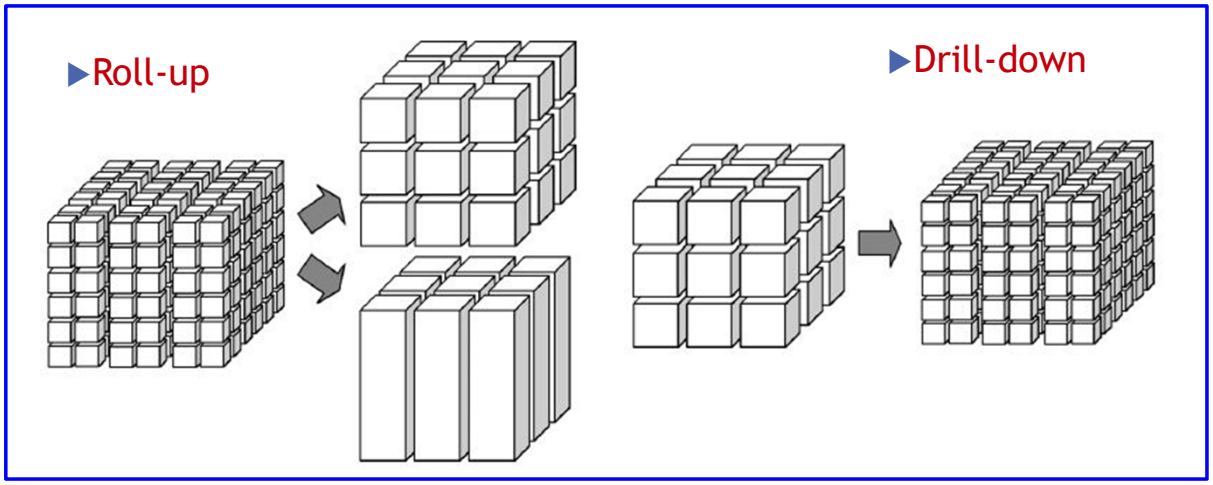
Analysis: Querying a data warehouse

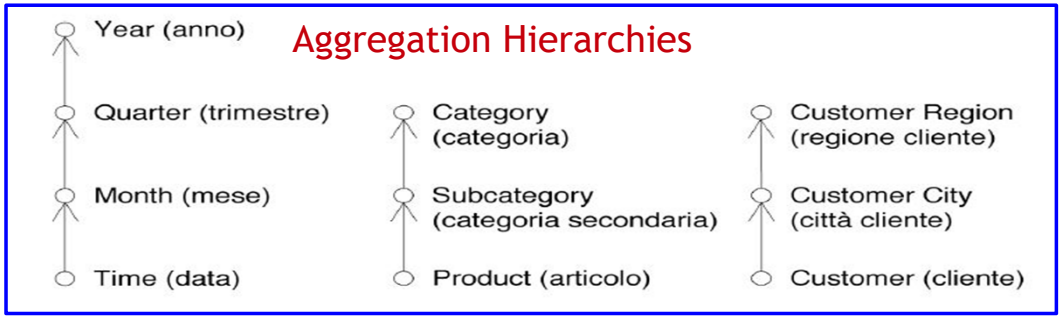
- **OLAP**
- An OLAP session consist of a navigation path that corresponds to an analysis process for fact according to different viewpoints and at different detail levels
- this path is turned into a sequence of queries
- the result of queries are multidimensional
- Each step of an analysis session is characterized by an OLAP operator that turns the latest query into a new one.
- Most common operators are:
  - **Roll-up / Drill-down**
  - **Slice-and-dice**
  - **Pivoting**
  - **Drill-across**
  - **Drill-through**

Roll-up and Drill-down

▶ **Roll-up**: increasing data aggregation; it removes a detail level from an hierarchy

▶**Drill-down (complement operation)**: reduces data aggregation and adds a detail level to the hierachy.





Roll-up

Roll-up on the temporal hierachy

Metrics / Customer Region / Month	Dollar Sales										
	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
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

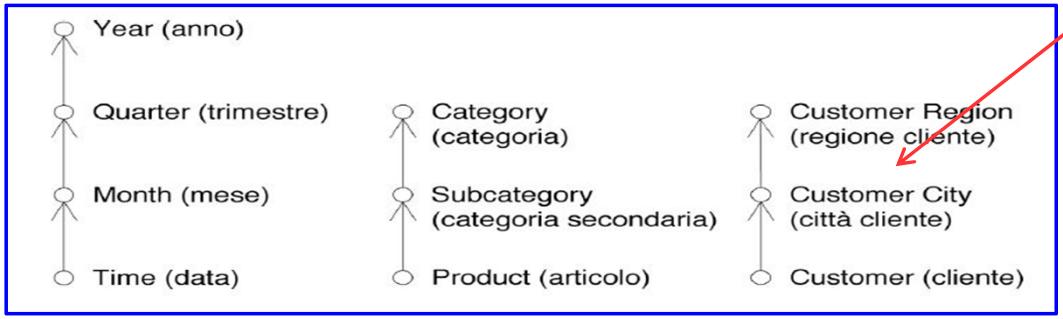
Metrics / Customer Region / Quarter	Dollar Sales										
	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany	Canada
Q1 1997	$ 1.526	$ 1.249	$ 978	$ 1.885	$ 3.650	$ 4.855	$ 1.834	$ 2.552	$ 1.920	$ 643	$ 663
Q2 1997	$ 2.979	$ 1.415	$ 2.664	$ 1.582	$ 1.130	$ 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

Drill-down

Drill-down on the customer dimension

Metrics	Dollar Sales										
Customer Region / Quarter	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

Metrics	Dollar Sales													
Customer City / Quarter	Arlin	San Pedro	Springfield	Chappel Hill	Scranburg	Pebble Beach	Martinsville	Maddon	Peoria	Pecos	Lake Barkley	Alcameda	Fingers Lake	S
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			

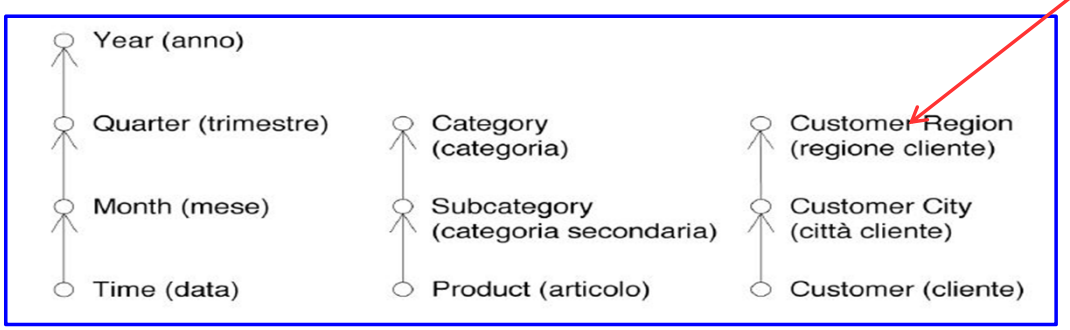


Drill-down

Drill-down by adding the Customer dimension

Category	Metrics: Dollar Sales — Year: 1997	1998
Electronics	$ 10.515	$ 29.299
Food	$ 5.303	$ 5.630
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

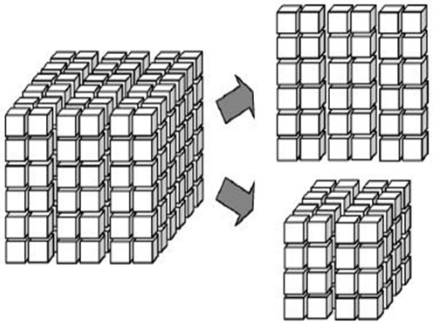
Category	Metrics: Dollar Sales — Customer Region: North East		Mid-Atlantic		South East		Central		South		North West	
	Year: 1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998
Electronics	$ 138	$ 1.184	$ 1.774	$ 4.529	$ 384	$ 1.892	$ 138	$ 7.232	$ 2.346	$ 651	$ 2.554	$ 9.488
Food	$ 759	$ 538	$ 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



Slicing and dicing

- **Slicing**: the operation reduces the dimensions of a cube by fixing the value of one or more dimensions.

- **Dicing**: the operation reduces a cube by expressing a selection expression for the values of one or more dimensions



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	$ 130	$ 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	$ 554	$ 386	$ 300	$ 978	$ 416	$ 48	$ 38	
	1998		$ 608	$ 559	$ 1.096	$ 611	$ 464	$ 316	$ 573	$ 257	$ 198	$

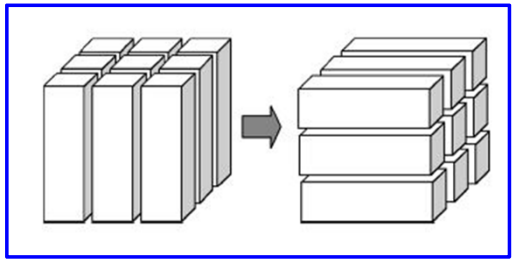
▶Year=1998

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	

Pivoting

▶**Pivoting**: implies a changing in layouts. It aims at analyzing an individual group of information from a different viewpoint.

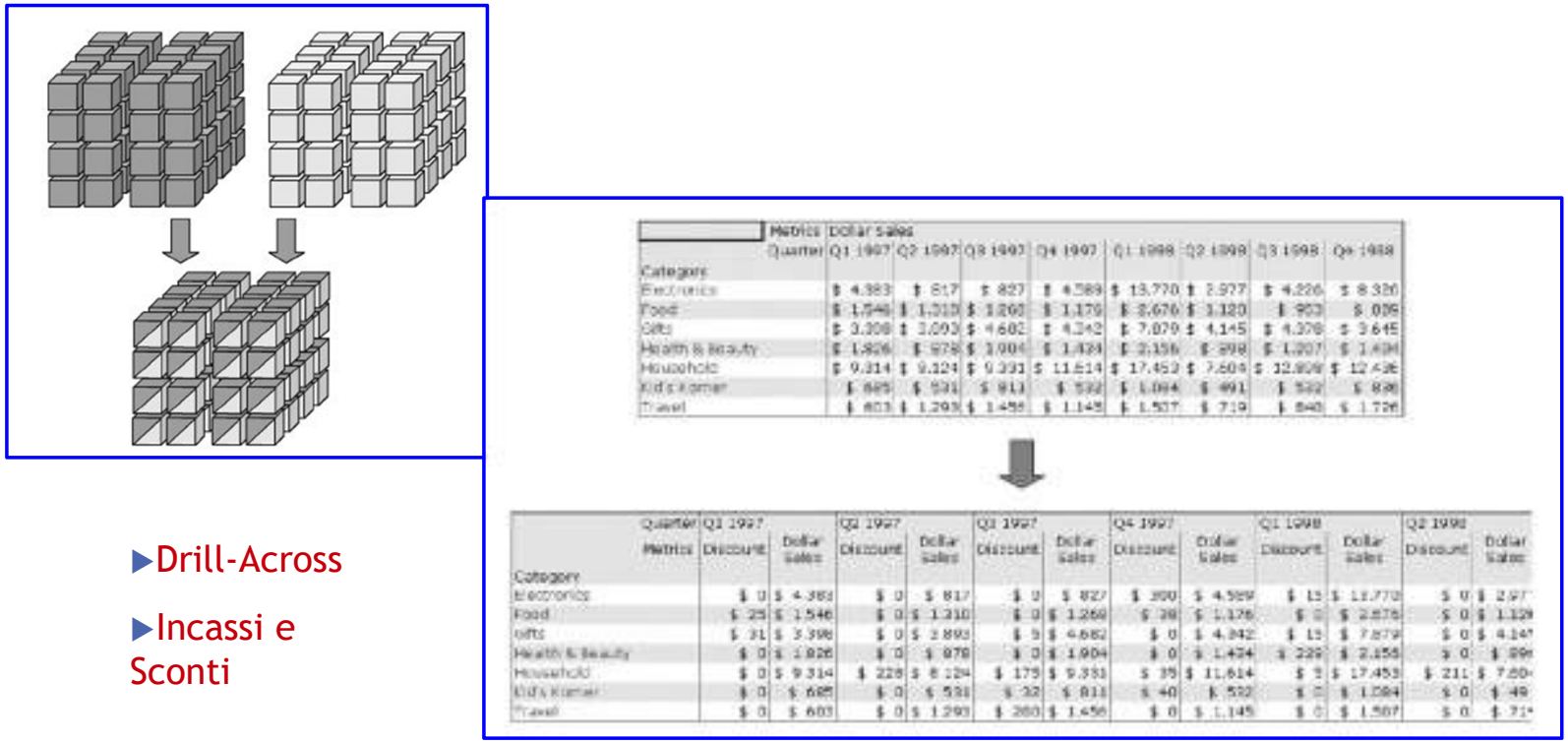


Category	Year	Metrics / Dollar Sales
Electronics	1997	$ 10.616
	1998	$ 29.299
Food	1997	$ 5.300
	1998	$ 5.638
Gifts	1997	$ 16.315
	1998	$ 20.047
Health & Beauty	1997	$ 6.042
	1998	$ 5.665
Household	1997	$ 38.383
	1998	$ 50.391
Kid's Korner	1997	$ 2.559
	1998	$ 2.943
Travel	1997	$ 4.497
	1998	$ 4.792

Category	Metrics / Year	Dollar Sales 1997	Dollar Sales 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

Drill-Across

- **Drill Across**: stands for the opportunity to create a link between two or more interrelated cubes in order to compare their data.
- **Drill through**: switches from multidimensional aggregate data in data marts to operational data in sources or in the reconciled layer



- Drill-Across
- Incassi e Sconti

Implementing a DW

- There are different approaches to implementing data warehouses. They are related to the logical model used to represent data:
- **ROLAP, Relational OLAP:** Implementation based on relational DBMSs.
- It takes advantage of the available corporate experience with relational database usage and management and the top performance and flexibility standards of relational DBMS
- Relational DBMSs do not include primitives for the concepts of dimension measure and hierarchy.
- Specific types of schemata must be created so to represent the multidimensional model in terms of basic relational elements (star schemata, snow-flake etc)
- The main problem results from the performance hit caused by costly joint operation between larger tables
- To improve performance:
  - **Denormalization**
  - **Fragmentation**
  - **Materialized Views**

Implementing a DW

- **MOLAP, Multidimensional OLAP**: Implementation based on Multidimensional DBMSs.
- Different from ROLAP a MOLAP system is based on ad hoc logical model that can be used to represent multidimensional data and operation directly
- The multidimensional database physical stores data in arrays and the access to it is positional
- The advantage is that multidimensional operations can be performed in easy way
- A criticism of MOLAP implementation is related to handling sparsity of data.
- **HOLAP, Hybrid OLAP**: Implementation based on both Multidimensional and Relational DBMSs.
- Aims at mixing the advantages of both basic solutions
- The largest amount of data should be stored in an RDBMS to avoid the problems caused by sparsity
- The multidimensional system stores only the information users most frequently need to process