POLITECNICO
MILANO 1863

Data Warehouse

Cinzia Cappiello
A.A. 2023-2024

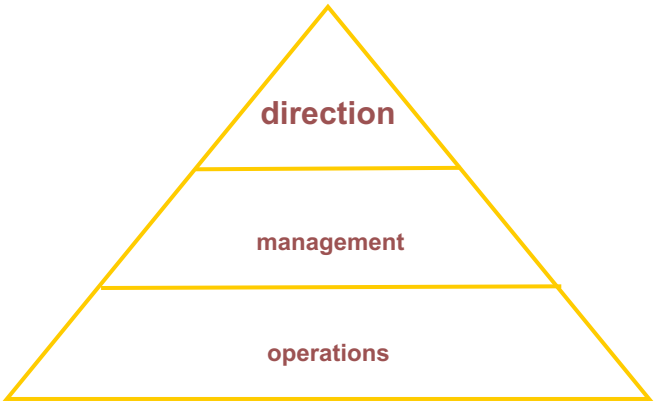
Outline

- What is a Data Warehouse?
- Data Warehouse Architecture
- Data Warehouse operations

Introduction

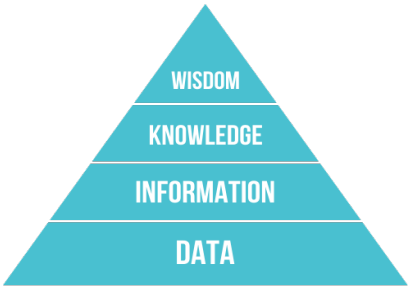
3

Business Processes' Pyramid

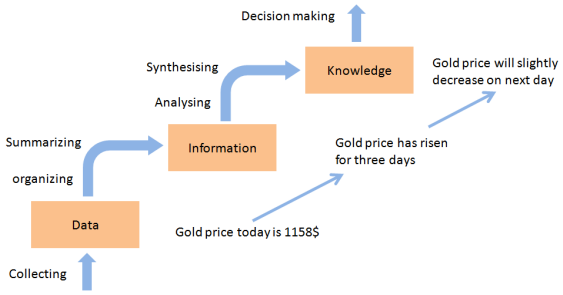


4

Knowledge pyramid



[wikipedia]



5

scope of decisions

	Operational		Tactical		Strategic
Accuracy	High	⟷			Low
Level of detail	Detailed	⟷			Aggregate
Time horizon	Present	⟷			Future
Frequency of use	High	⟷			Low
Source	Internal	⟷			External
Scope of information	Quantitative	⟷			Qualitative
Nature of information	Narrow	⟷			Wide
Age of information	Present	⟷			Past

Carlo Vercellis 2006

6

OLTP and OLAP systems

OLTP (Standard DB)	OLAP
Mostly updates	Mostly reads
Many small transactions	Queries are long and complex
Current snapshot	History
Raw data	Summarized, reconciled data
Thousands of users (e.g., clerical users)	Hundreds of users

7

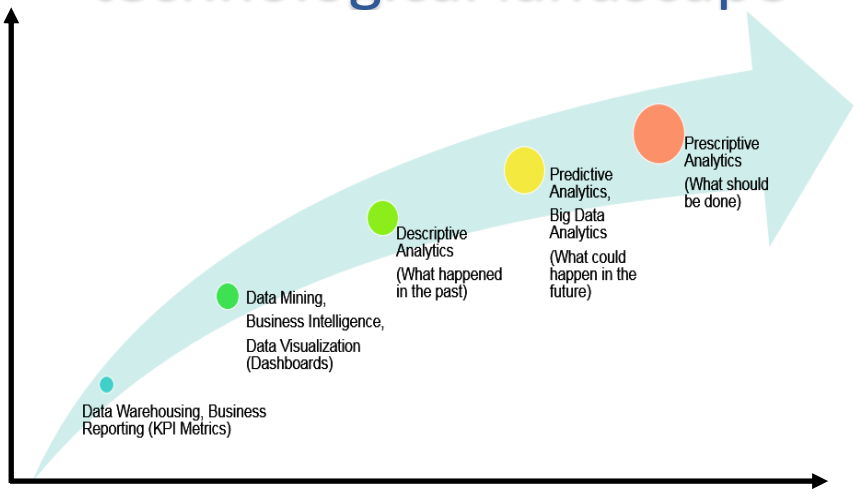
OLAP properties

- FASMI
  - Fast
  - Analytical
  - Shared
  - Multidimensional
  - Informational

8

8

Data analysis – the technological landscape



9

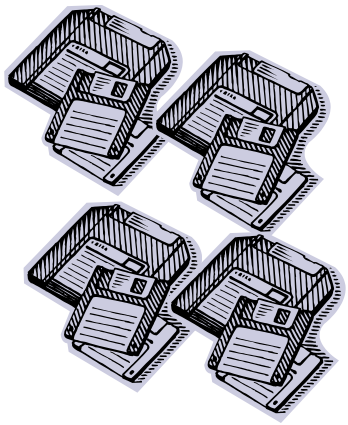
Data Warehouse

10

What is a Data Warehouse?

- Data should be **integrated across the enterprise(s)**
- **Summary data** provide real value to the organization
- **Historical data** hold the key to understanding data over time
- What-if capabilities are required

What is a Data Warehouse?



A single, complete and consistent store of data obtained from a variety of different sources made available to end users, *so that they can understand and use it in a business context.*

[Barry Devlin]

An alternative definition of Data Warehouse

Data

A data warehouse is a process for *transforming data into information* and for making it available to users *in a timely enough manner to make a difference*.

[Forrester Research, April' 96]

13

Data Warehouse

As a dataset: decision support database maintained separately from the organization's operational database

As a process: technique for assembling and managing data from various sources with the purpose of answering business questions. Thus making decisions that were not previously possible

14

14

OLAP Properties

- OLAP systems are characterized by FASMI properties:
  - Fast
  - Analytical
  - Shared
  - Multidimensional
  - Informational

15

Data Warehouse

- A Data Warehouse is a
  - subject-oriented: “the data contained in a data warehouse are primarily concerned with the main entities of interest for the analysis, such as products, customers, orders and sales”
  - Integrated: “The data originating from the different sources are integrated and homogenized as they are loaded into a data warehouse”
  - time-variant: “All data entered in a data warehouse are labelled with the time period to which they refer”
  - non-volatile (persisent): “Once they have been loaded into a data warehouse, data are usually not modified further and are held permanently “
- collection of data that is used primarily in organizational decision making.

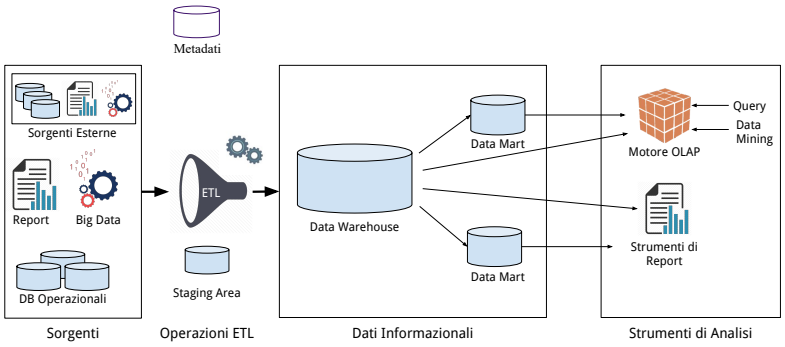
[Bill Inmon, Building the Data Warehouse, 1996]

16

Dimensions of a Data Warehouse

- Data warehouses are **very** large databases
  - Terabytes (10^{12} bytes)
  - Petabytes (10^{15} bytes): e.g. Geographic Information Systems
  - Exabytes (10^{18} bytes): e.g. National Medical Records
  - Zettabytes (10^{21} bytes): e.g. Weather reports, including images
  - Yottabytes (10^{24} bytes): e.g. Intelligence Agency Videos

Data Warehouse - architecture



Where is a DW useful

- **Commerce**: sales and complaints analysis, client fidelization, shipping and stock control
- **Manufacturing plants**: production cost control, provision and order support
- **Financial services**: risk and credit card analysis, fraud detection
- **Telecommunications**: call flow analysis, subscribers' profiles
- **Healthcare structures**: patients' ingoing and outgoing flows, cost analysis

19

ETL (Extraction, Transformation, Loading)

- **Extraction**: data are extracted from the available internal and external sources.
- **Transformation**: the goal of the transformation phase is to improve the quality of the data extracted. Some of the main operations that are executed are:
  - Data Cleaning
  - Reconciliation, Entity Matching
  - Data standardization
  - Deduplication
- **Loading**: after being extracted and transformed data are loaded into the data warehoise

21

Metadata

- Metadata contain the following data:
  - Information about the data warehouse structure (e.g., dimensions, hierarchies, fact)
  - Information about values stored in the data warehouse: each attribute is characterized by its provenance, e.g., which is the data sources from which data were extracted and the transformations to which they have been subjected
  - Usage statistics of the data warehouse, e.g. number of accesses to a field
  - Description of the application domain and related data properties, data ownership and loading policies

Examples of data warehouse queries

- Show total sales across all products at increasing aggregation levels for a geography dimension, from state to country to region, for 2017 and 2018.
- Create a cross-tabular analysis of our operations showing expenses by territory in South America for 2017 and 2018. Include all possible subtotals.
- List the top 10 sales representatives in Asia according to sales revenue for automotive products in year 2018, and rank their commissions.

OLAP-oriented data models

- must support sophisticated analyses and computations over different dimensions and hierarchies
- Must guarantee fast responce time even to complex queries
- Most appropriate data model: multidimensional model

24

Dimensional Fact Model

- Allows one to describe a set of

  fact schemata

- The components of a fact schema are:
  - Facts
  - Measures
  - Dimensions
  - Dimension Hierarchy

25

Dimensional Fact Model

- A **fact** is a concept that is relevant for the decisional process; typically it models a set of events of the organization
- A **measure** is a numerical property of a fact
- A **dimension** is a fact property defined w.r.t. a finite domain; it describes an analysis coordinate for the fact, it is a perspective for analysing data
- **Dimension Hierarchy**: relate low-level (detailed) concepts to higher-level (general concepts)
  - Example: Store – City – Region/Province – Country

26

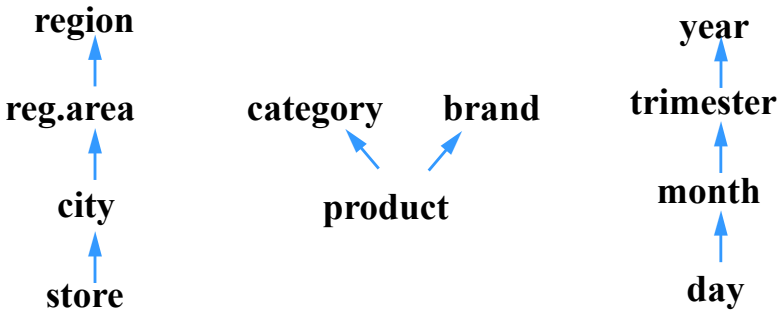
Dimensional Fact Model

- The multidimensional view of data is represented as a data cube or an hypercube
- Cube dimensions are the search keys
- Each dimension may be hierarchical
  - DATE {DAY-MONTH-TRIMESTER-YEAR}
  - PRODUCT {BRAND - TYPE - CATEGORY}

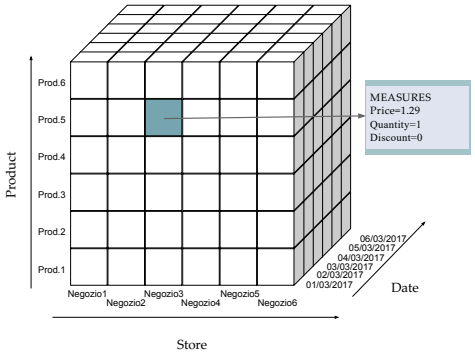
    (e.g. LAND ROVER - CARS - VEHICLES)
- Cube cells contain metric values

27

Dimensions and hierarchies



Example sales



* ESSELUNGA S.p.A. *
VIA CORRIDONI n.30 - BERGAMO (BG)
P.I.: 04916380159

	EURO
GSK DENT. AQUAFRESH	1,19
GSK DENT. AQUAFRESH	0,99
GSK DENT. AQUAFRESH	0,99
GSK DENT. AQUAFRESH	0,99
GSK DENT. AQUAFRESH	0,99
SCHIACCIATE OLIV&MARI	1,29
CRACKERS OLIVIA&MARIN	1,79
GRISS POMODORO OL&MAR	1,49
8 x 2,99	
BONDUE. COCCOLE SPINA	23,92
SCONTO FIDATY 30%	7,20-S

sconTomaggio.com

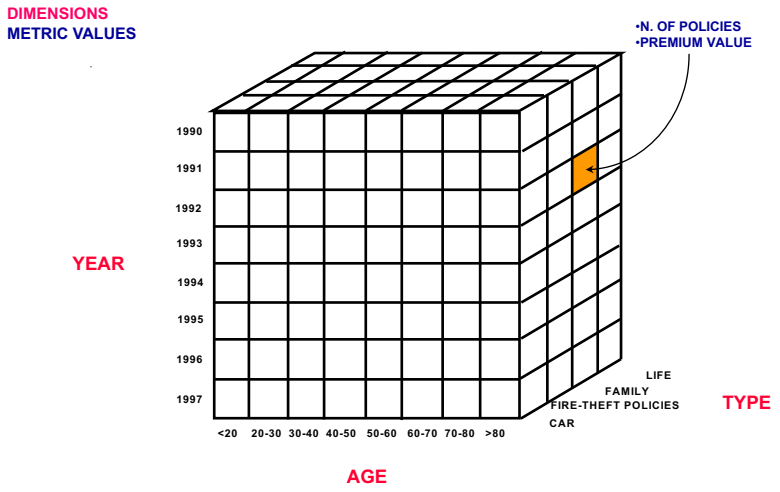
TOTALE EURO	26,44 *
PAGAMENTO SCONTO EURO	0,04
PAGAMENTO BUONI SCONTO	9,50
PAGAMENTO BANCOMAT	16,90
RESTO	0,00 *

N. CARTA FIDATY: 040********92

---------- NUOVA RACCOLTA PUNTI ----------

SALDO AD OGGI PUNTI	4.744
PUNTI SULLA SPESA	42
TOT.PUNTI FIDATY	42
NUOVO SALDO PUNTI	4.786

Example: An Insurance Company Data Cube



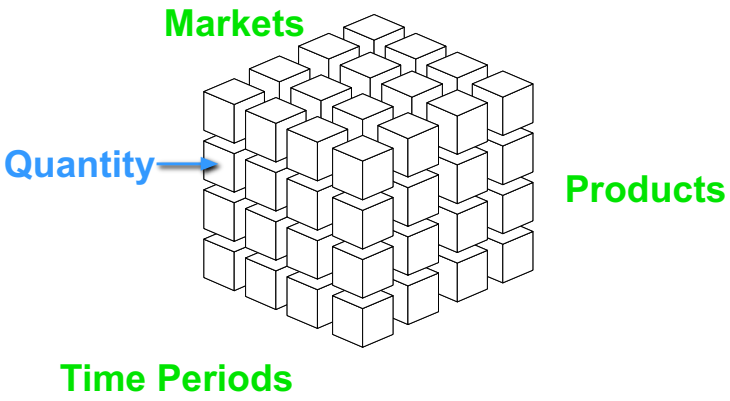
30

Examples

- Store chain
  - Fact: sales
  - Measures: sold quantity, gross income
  - Dimensions: product, time, zone
- Telecom Operator
  - Fact: phone call
  - Measures : cost, duration
  - Dimensions: caller subscriber, called subscriber, time

31

Multidimensional Representation



Sales Cube

32

OLAP operations

- Slice/Dice
- Roll up/Drill down

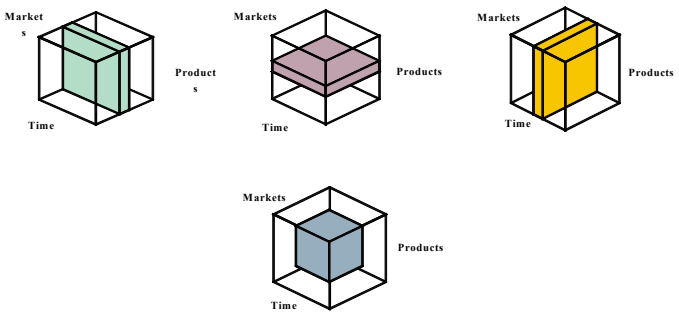
33

Typical Olap Operations

- **slice-and-dice**
  - The **slice** operation performs a selection on one dimension of the given cube, resulting in a subcube
  - The **dice** operation defines a subcube by performing a selection on two or more dimension
- **roll-up**
  - Aggregates data at a higher level – e.g. last year's sales volume per product category and per region
- **drill-down**
  - De-aggregates data at the lower level – e.g. for a given product category and a given region, show daily sales
- **pivoting**
  - Selects two dimensions to re-aggregate data (cube re-orientation)
- **ranking**
  - Sorts data according to predefined criteria
- traditional operations (select, project, join, derived attributes, etc.)

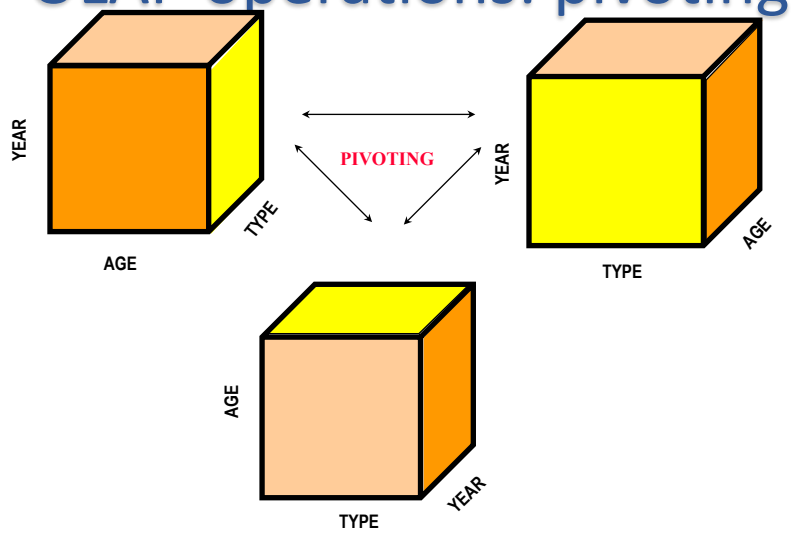
34

Slice/Dice operations



35

OLAP operations: pivoting



36

Pivoting

Fact table view:

sale	prodId	storeId	date	amt
	p1	c1	1	12
	p2	c1	1	11
	p1	c3	1	50
	p2	c2	1	8
	p1	c1	2	44
	p1	c2	2	4

Multi-dimensional cube:

day 2

	c1	c2	c3
p1	44	4	

day 1

	c1	c2	c3
p1	12		50
p2	11	8	

	c1	c2	c3
p1	56	4	50
p2	11	8	

Hector Garcia Molina, Data warehouse and OLAP

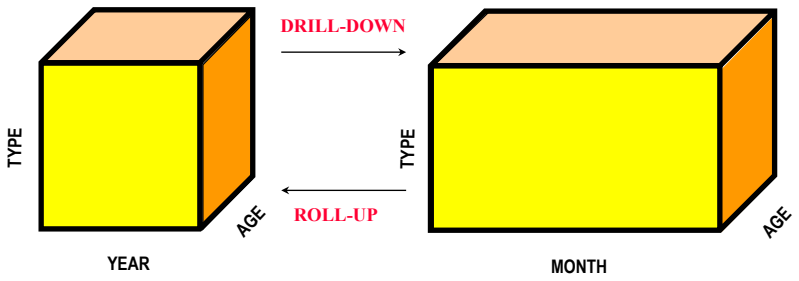
37

Category	Metrics / Year	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	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

38

OLAP OPERATIONS



39

Time

ID	Day	Month	Trimester	Year
T1	31/3/2018	March	1	2018
T2	02/4/2018	April	2	2018

Product

ID	Brand	Type	Category
P1	M1	Milk	Food
P2	M2	Bread	Food

Sales

Time	Product	Store	Price
T1	P1	PV1	3,5
T1	P2	PV2	4
T1	P1	PV3	4
T2	P2	PV3	5
T2	P2	PV1	4
T2	P1	PV2	3

Store

ID	Address	City	Region	State
PV1	Via Roma 1	Milan	Lombardia	Italy
PV2	Via Milano 4	Rome	Lazio	Italy
PV3	Via Torino 5	Milan	Lombardia	Italy

40

Sales Roll Up

Time	Product	Store	Price
T1	P1	PV1	3,5
T1	P2	PV2	4
T1	P1	PV3	4
T2	P2	PV3	5
T2	P2	PV1	4
T2	P1	PV2	3

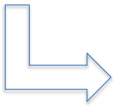
T1	P1	Lombardia	7,5
T1	P2	Lazio	4
T2	P2	Lombardia	9
T2	P1	Lazio	3

Amount of sales per region

41

Roll up

T1	P1	Lombardia	7,5
T1	P2	Lazio	4
T2	P2	Lombardia	9
T2	P1	Lazio	3



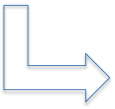
2018	P1	10,5
2018	P2	13

Amount of sales per year and product without considering the store

42

Roll up

2018	P1	10,5
2018	P2	13



2018	23,5

Amount of sales per year

43

Drill down

2018	23,5

2018	P1	10,5
2018	P2	13

Amount of sales per year and product

Drill down

2018	P1	10,5
2018	P2	13

1/2018	P1	7,5
1/2018	P2	4
2/2018	P1	3
2/2018	P2	9

Amount of sales per trimester

Roll-up

Month	Metrics / Customer Region	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germany
	Dollar Sales										
Jan 97		$ 620	$ 753	$ 30	$ 660	$ 2.405	$ 1.312	$ 440	$ 1.002	$ 1.002	$ 383
Feb 97		$ 258	$ 252	$ 800	$ 975	$ 160	$ 582	$ 744	$ 310	$ 799	$ 118
Mar 97		$ 648	$ 244	$ 148	$ 250	$ 1.085	$ 2.961	$ 650	$ 1.240	$ 119	$ 142
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
Jun 97		$ 842	$ 582	$ 1.281	$ 937	$ 240	$ 774	$ 176	$ 1.139	$ 652	$ 254
Jul 97		$ 652	$ 690	$ 486	$ 1.293	$ 605	$ 303	$ 818	$ 103	$ 124	$ 173
Aug 97		$ 1.783	$ 304	$ 1.032	$ 170	$ 398	$ 356	$ 432	$ 190	$ 241	$ 407
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
Nov 97		$ 39	$ 1.602	$ 1.082	$ 1.187	$ 842	$ 759	$ 745	$ 232	$ 101	$ 1.037
Dec 97		$ 381	$ 1.588	$ 343	$ 118	$ 1.459	$ 635	$ 2.021	$ 259	$ 210	$ 119
Jan 98		$ 311	$ 1.174	$ 2.634	$ 3.130	$ 954	$ 2.083	$ 1.351	$ 747	$ 426	$ 447
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
Apr 98		$ 407	$ 841	$ 524	$ 712	$ 133	$ 2.486	$ 49	$ 390	$ 1.298	$ 221
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
Jul 98		$ 586	$ 1.897	$ 412	$ 226	$ 406	$ 361	$ 1.628	$ 267	$ 1.011	$ 41
Aug 98		$ 894	$ 326	$ 792	$ 1.832	$ 1.199	$ 295	$ 1.816	$ 277	$ 102	$ 118
Sep 98		$ 338	$ 3.179	$ 505	$ 427	$ 99	$ 2.976	$ 885	$ 135	$ 85	$ 1.110
Oct 98		$ 544	$ 413	$ 1.467	$ 209	$ 679	$ 706	$ 556	$ 480	$ 485	$ 99
Nov 98		$ 671	$ 459	$ 1.471	$ 2.066	$ 701	$ 716	$ 986	$ 1.127	$ 154	$ 440
Dec 98		$ 836	$ 2.096	$ 1.726	$ 3.642	$ 395	$ 1.740	$ 1.943	$ 1.143	$ 366	$ 307

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

46

Roll-up

Category	Year	Metrics / Customer Region	North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France
		Dollar Sales									
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
Food	1997		$ 759	$ 682	$ 729	$ 262	$ 588	$ 469	$ 807	$ 156	$ 615
	1998		$ 530	$ 925	$ 959	$ 677	$ 213	$ 1.503	$ 261	$ 165	$ 175
Gifts	1997		$ 2.532	$ 1.355	$ 1.854	$ 1.413	$ 2.535	$ 2.132	$ 1.904	$ 908	$ 375
	1998		$ 1.955	$ 2.785	$ 2.800	$ 2.695	$ 1.813	$ 2.844	$ 1.778	$ 1.158	$ 717
Health & Beauty	1997		$ 624	$ 640	$ 1.317	$ 647	$ 588	$ 754	$ 654	$ 143	$ 292
	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
	1998		$ 5.787	$ 5.320	$ 5.416	$ 6.812	$ 4.334	$ 5.008	$ 7.588	$ 2.139	$ 3.649
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		$ 2.550
	1998		$ 2.943
Travel	1997		$ 4.497
	1998		$ 4.792

47

Drill-down

Metrics	Dollar Sales	
Year	1997	1998
Category		
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 Corner	$ 2.559	$ 2.943
Travel	$ 4.497	$ 4.792

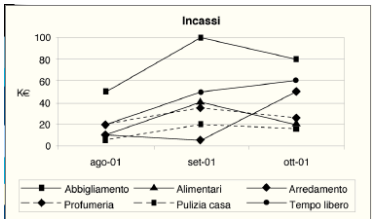
Metrics	Dollar Sales											
Customer Region	North-East		Mid-Atlantic		South-East		Central		South		North-West	
Year	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998	1997	1998
Category												
Electronics	$ 138	$ 1.184	$ 1.774	$ 4.529	$ 384	$ 1.892	$ 138	$ 7.232	$ 2.346	$ 651	$ 2.554	$ 9.488
Food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

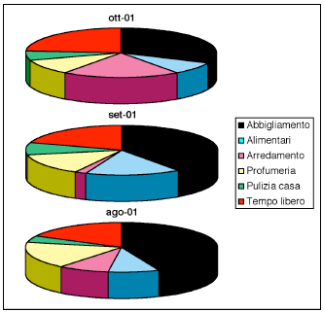
48

Visualization and Reports

- Data may be visualized graphically, in an Excel-like format: tables, hystograms, graphics, 3D surfaces, etc.

incassi (K€)	Ottobre 2001	Settembre 2001	Agosto 2001
Abbigliamento	80	100	50
Alimentari	20	40	10
Arredamento	50	5	10
Profumeria	25	35	20
Pulizia casa	15	20	5
Tempo libero	60	50	20





51

Aggregate Queries

Examples:

- Total sales per product category, per supermarket, per day
- Total monthly sales for all the products, per supermarket
- Total monthly sales per category per supermarket
- Avg. monthly sales per category, for all supermarkets