

BI 1.0 & Data Warehousing

The turbulent evolution of data analysis

Matteo Francia <m.francia@unibo.it>



Glossary

Database

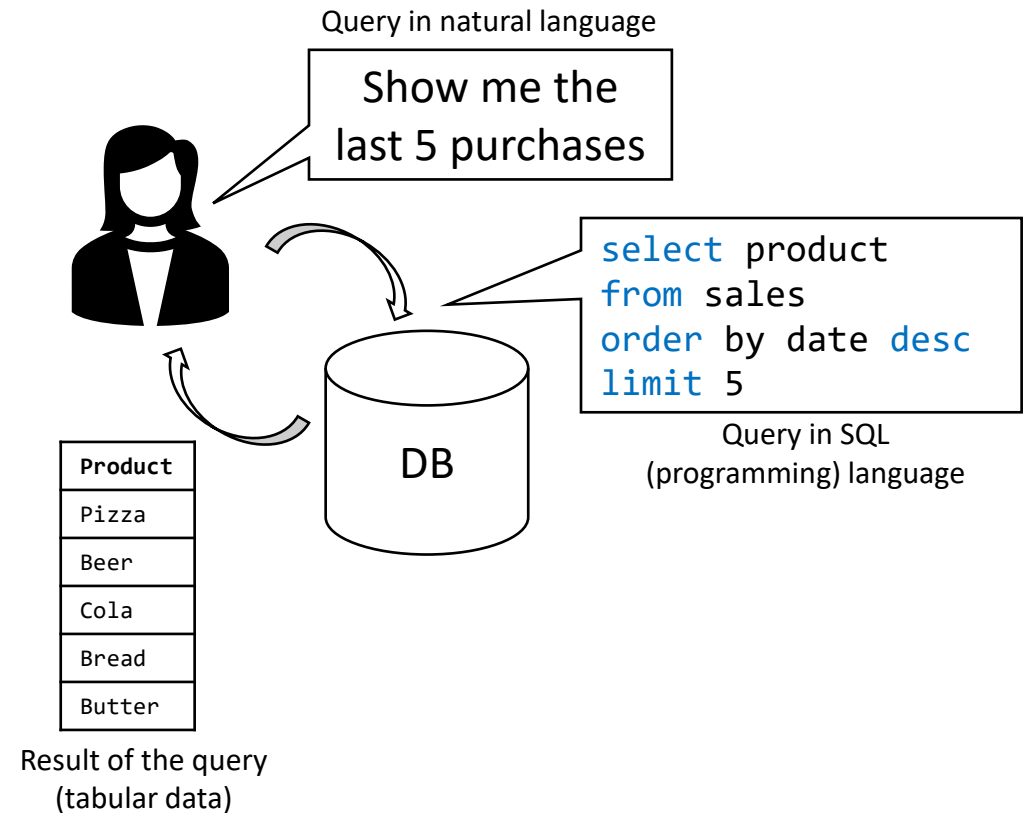
- Any collection of data that is specially organized for rapid search and retrieval by a computer
- Relational databases organize data as tables

Operational database

- A system that efficiently manages and processes real-time data for day-to-day business operations

Query

- A *database query* can be defined as a request for data from a database
- Usually, it is expressed in a structured query language such as SQL



Query workloads

Imagine a supermarket.

OLTP (Online Transaction Processing) is like the cash register where you buy things.

- When you pick up some milk and go to the checkout, the system records your purchase.
- It's all about day-to-day operations — sales, refunds, payments.
- OLTP is fast because it needs to handle many people buying things at the same time.

OLAP (Online Analytical Processing) is like the manager's office, where they analyze all purchases.

- The manager looks at data from many days to answer questions like:
 - What products are selling best?
 - What times are busiest?
 - Should we order more apples next week?
- OLAP is slower, but it's okay because it's about studying and making decisions, not quick sales.

Query workloads

Key idea:

- OLTP = Quickly saving and updating information when people do everyday activities.
- OLAP = Looking at a lot of information to find patterns and make smart decisions.

In short:

- OLTP = Doing everyday work (fast transactions)
- OLAP = Studying the big picture (deep analysis)

Examples of OLTP queries

- A retailer uses a point-of-sale system to complete transactions online and in-store. OLTP system processes each transaction and creates a database of the information for each transaction. It then sends a request to the customer's credit card company, which approves the charge to the card. The system records the payment and deducts one item from the store's inventory. This transaction follows all the rules of a successful OLTP system and helps the business process payments quicker and manage its inventory more efficiently.
- A couple has a joint account at their bank, and each person attempts to withdraw the full balance of \$4,502.34 on the same day at the same time. The OLTP system fails the transactions for violating the concurrence rule. The couple has to withdraw money from the same bank account at separate times for the transaction to process successfully. Otherwise, the bank might provide the full balance for each party, creating a deficit of \$4,502.34 in the account.

Examples of OLAP queries

- “What is the relationship between performance of shares of PC manufacturers and quarterly profits along the last 5 years?”
- “What are the types of orders that will maximize revenues?”

ICT in companies

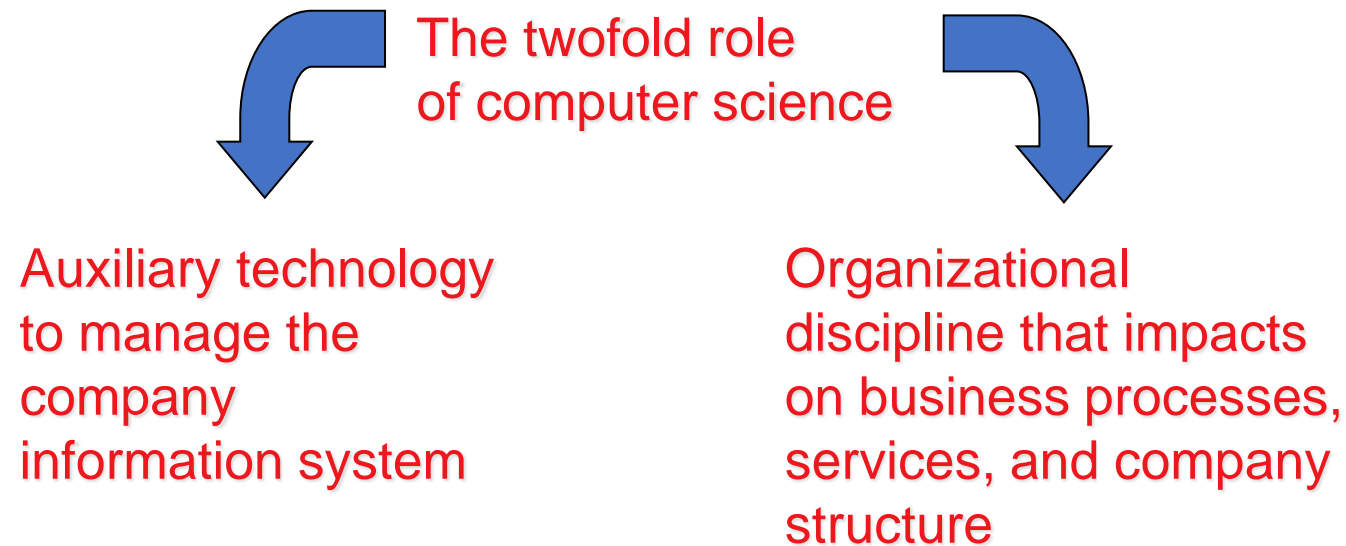
Up to some years ago, the main goal of databases in companies has been that of storing **operational data**, i.e., data generated by operations carried out within business processes

Computer science was seen as a **subsidiary discipline** that makes information management faster and cheaper, but does not create profits in itself



The evolution of information systems

- The role of computer science in companies has radically changed since the early 70's. ICT systems turned from simple **tools to improve process efficiency** into **key factors of company organizations** capable of deeply impacting on the structure of business processes



The new role of computer science in decision making

An **exponential increase** in operational data has made computers the only tools suitable for providing data for decision-making performed by business managers

The massive use of techniques for analyzing enterprise data made information systems a **key factor to achieve business goals**



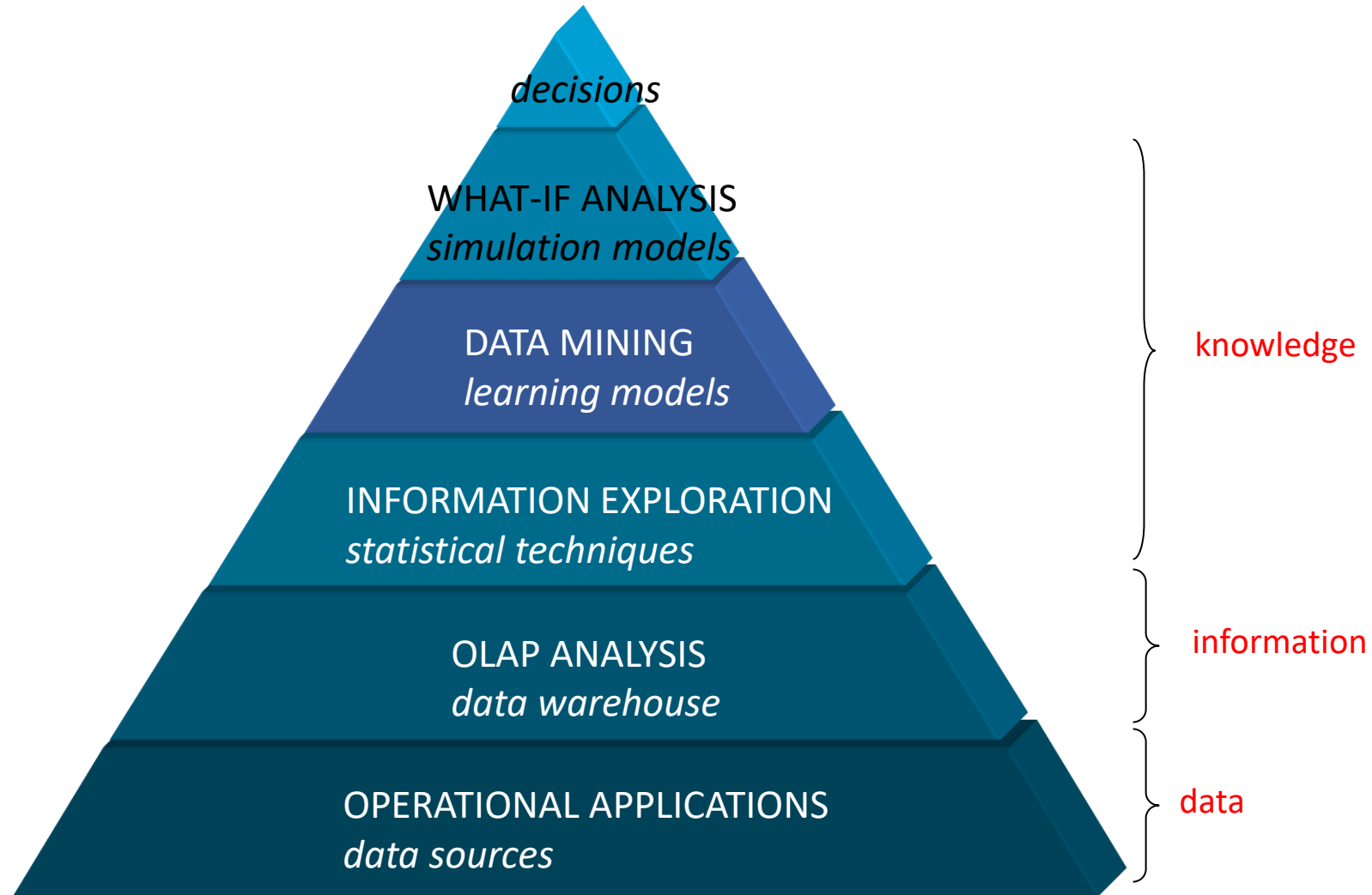
Business intelligence

A set of tools and techniques that enable a company to transform its business data into timely and accurate **information for the decisional process**

- Business intelligence systems are used by decision makers to **get a comprehensive knowledge of the business** and of the factors that affect it, as well as to **define and support their business strategies**
- The goal is to enable data-based decisions aimed at gaining competitive advantage, improving operative performance, responding more quickly to changes, increasing profitability and, in general, **creating added value for the company**



The BI1.0 pyramid



Data Warehousing

From data to information

Information assets are immensely valuable to any enterprise, and because of this, these assets must be properly stored and readily accessible when they are needed

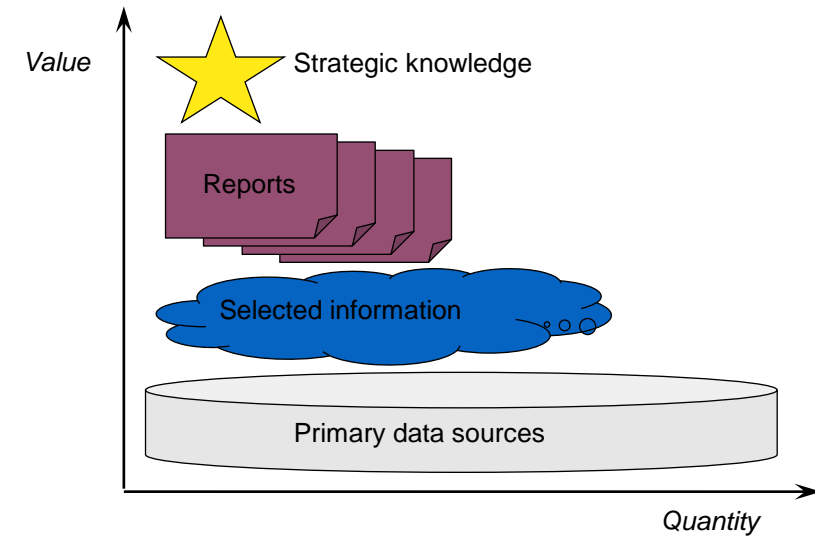
However, the availability of too much data makes the extraction of the most important information difficult, if not impossible

~~data = information~~

From data to information

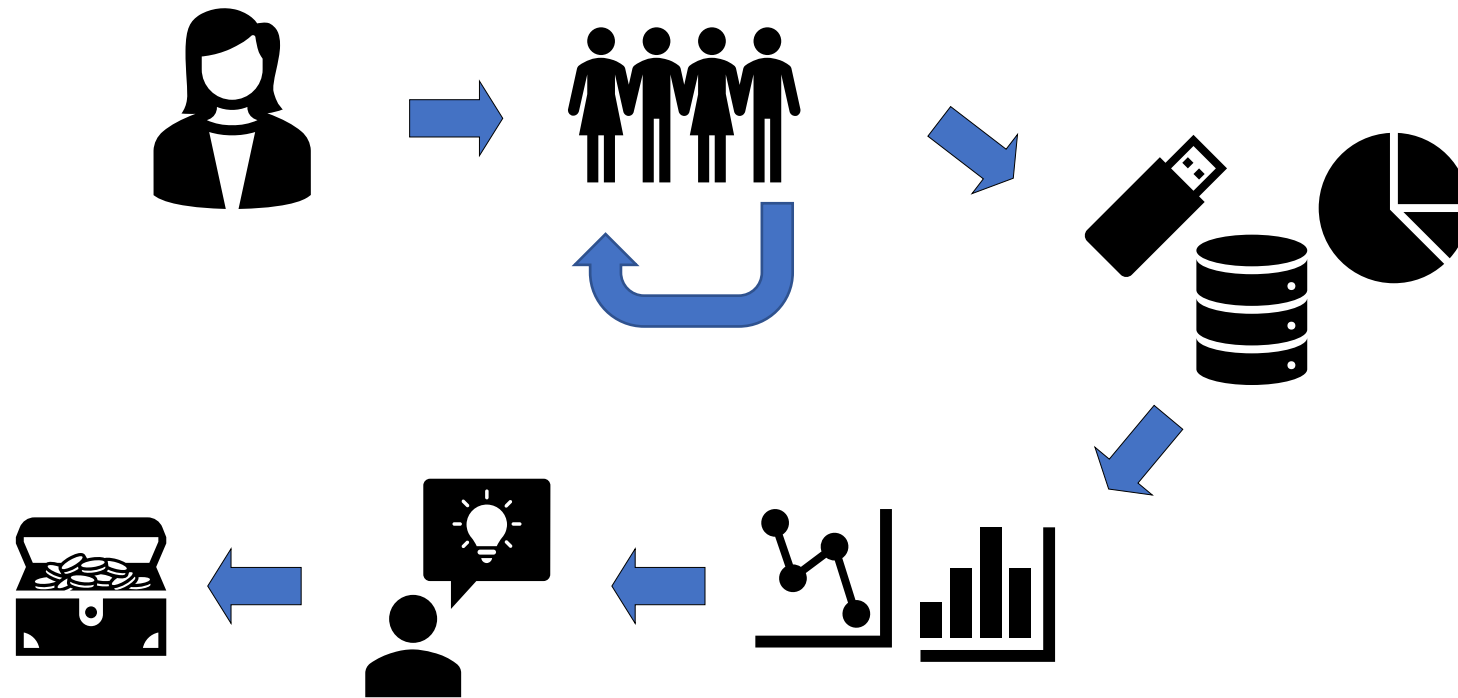
A company must have quick access to all the information needed for decision making

Strategic knowledge is extracted from the huge amount of operational data stored in enterprise databases, through a selection and aggregation process



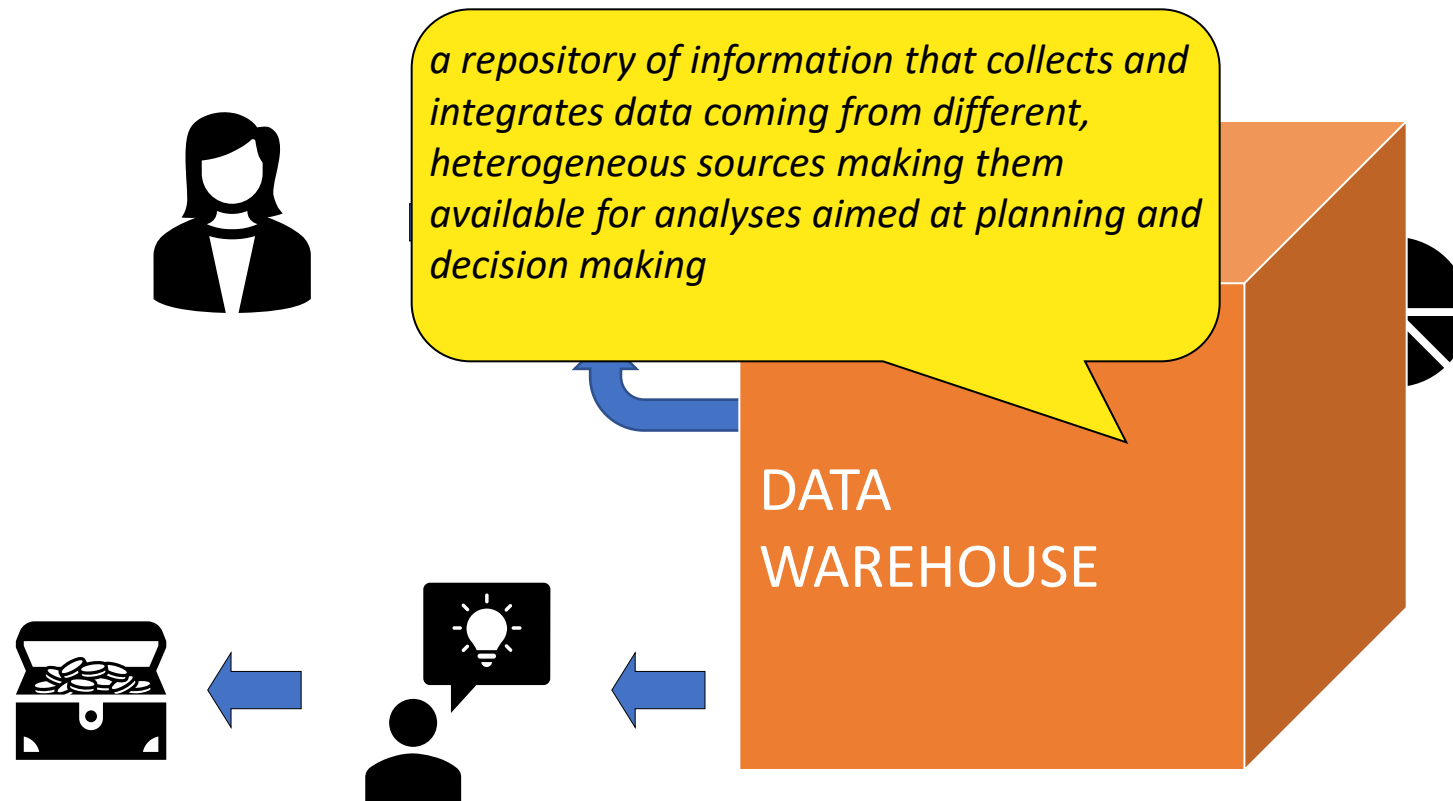
A typical scenario...

... is that of a large company, with several branches, whose managers wish to quantify and evaluate the contribution given from each branch to the global profit



A typical scenario...

... is that of a large company, with several branches, whose managers wish to quantify and evaluate the contribution given from each branch to the global profit



Queries

OLTP:

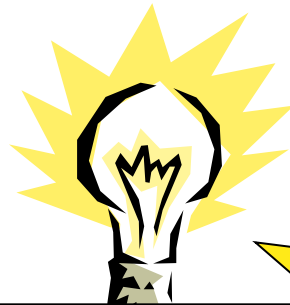
- They execute transactions that generally read/write a small number of tuples from/to many tables connected by simple relations
- The essential workload core is “frozen” in application programs, and ad hoc data queries are occasionally run for data maintenance

OLAP:

- Dynamic, multidimensional analyses that need to scan a huge amount of records to process a set of numeric data summing up the performance of an enterprise
- Interactivity is an essential property for analysis sessions, so the actual workload constantly changes as time goes by

OLTP and OLAP

A mix of analytical queries with transactional routine queries inevitably slows down the system, and this does not meet the needs of users of both types of queries



separate *online analytical processing (OLAP)* from *online transactional processing (OLTP)* by creating a new repository that integrates data from various sources and then makes data available for analysis and evaluation aimed at decision-making processes

Data Warehousing:

A collection of methods, techniques, and tools used to support *knowledge workers* (senior managers, directors, managers, and analysts) to conduct data analyses that help with performing decision-making processes and improving information resources

User claims

- ✎ *We have heaps of data, but we cannot access them!*
- ✎ *How can people playing the same role achieve substantially different results?*
- ✎ *We want to select, group, and manipulate data in every possible way!*
- ✎ *Show me just what matters!*
- ✎ *Everyone knows that some data is wrong!*

R. Kimball, The Data Warehouse Toolkit



Requirements for the warehousing process

Accessibility to users not very familiar with IT and data structures

Integration of 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

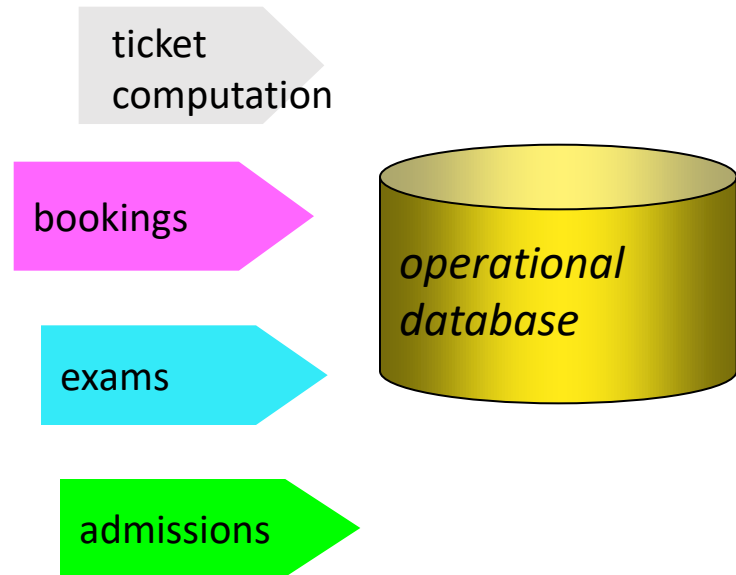
The Data Warehouse

In the middle of this process, a data warehouse is a data repository that fulfills the requirements

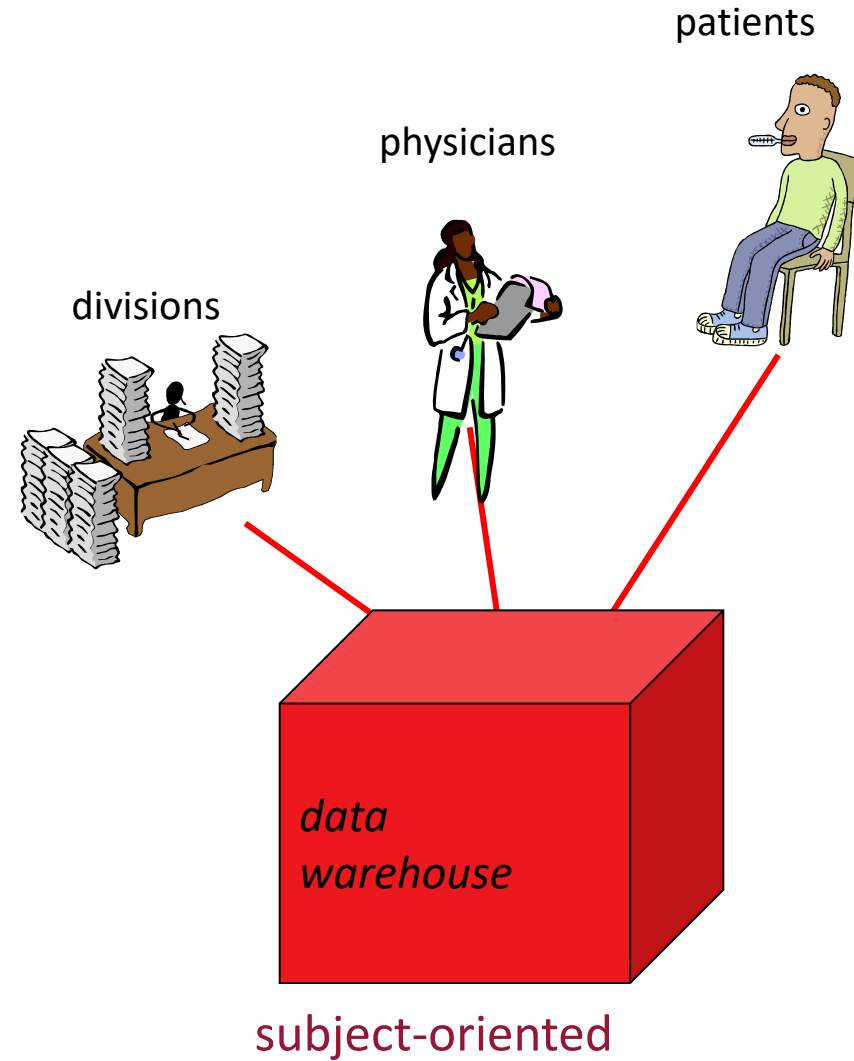
➤ A *data warehouse* is a collection of data that supports decision-making processes. It provides the following features:

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

...subject-oriented



application-oriented



...integrated and consistent

Data warehouses take advantage of multiple data sources, such as data extracted from production and then stored to enterprise databases, or even data from a third party's information systems. A data warehouse should provide a unified view of all the data.



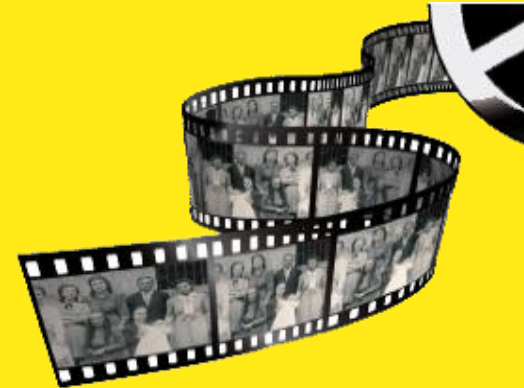
...shows its temporal evolution

Operational DB



Limited historical content,
time is often not part of the
keys, data are updated

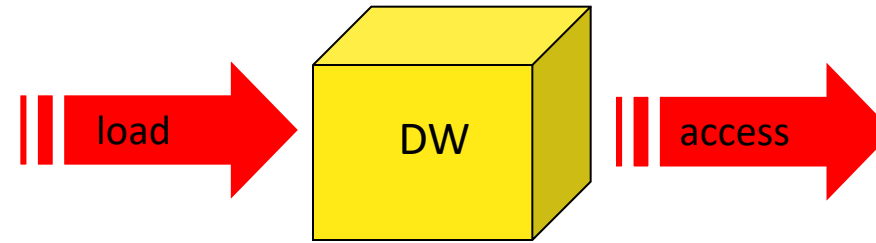
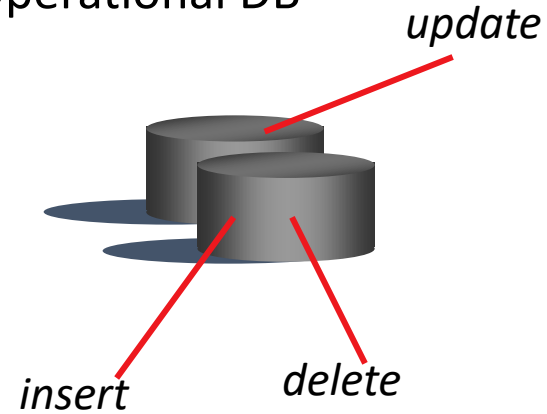
DW



Rich historical content,
time is part of the keys,
a snapshot of data taken at a
given time cannot be updated

...non volatile

Operational DB



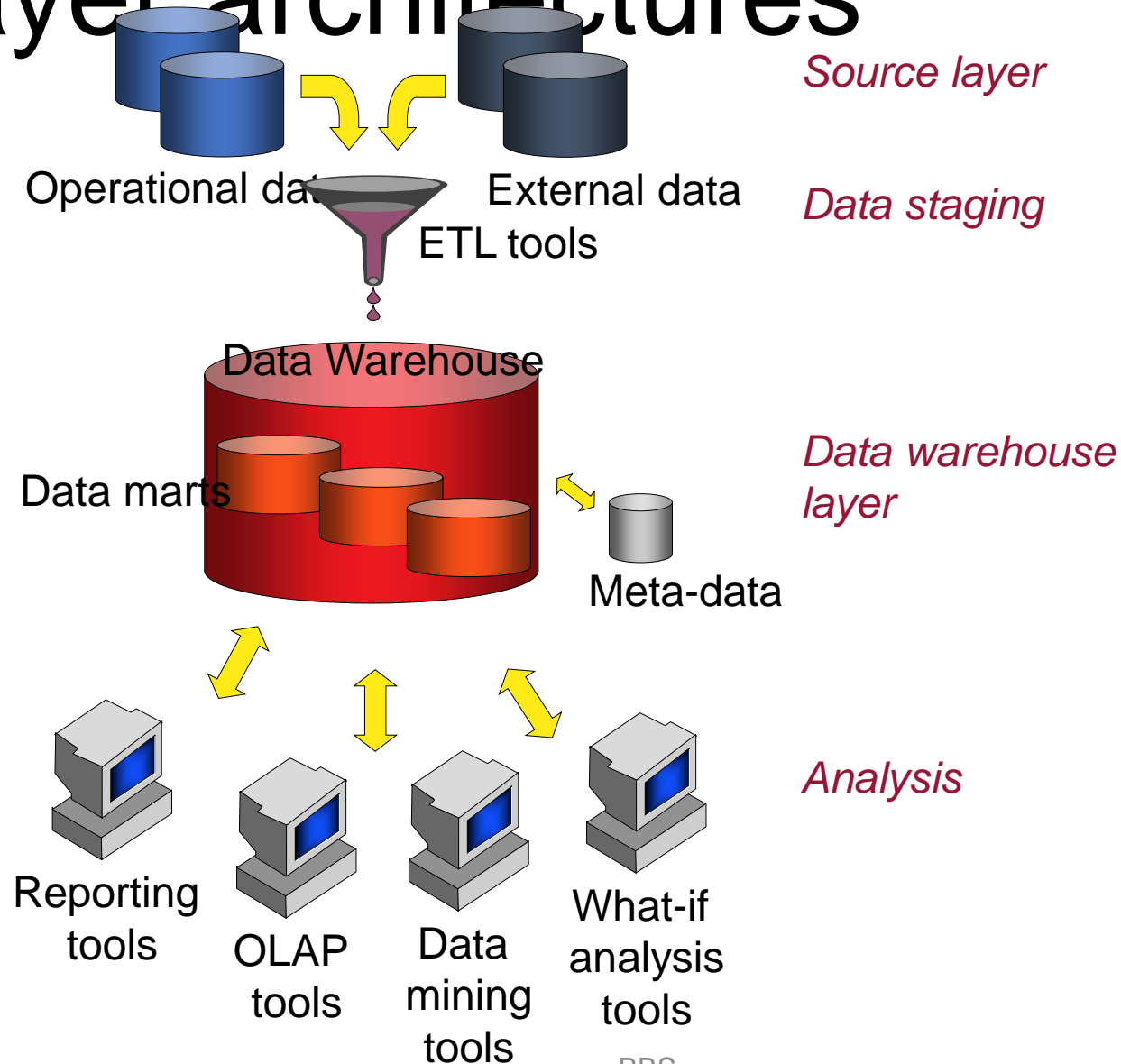
Huge data volumes:
from 50 GBs to some TBs
in a few years

- no need for advanced transaction management techniques required by operational applications
- key problems are query-throughput and resilience

Summarizing

	Operational DBs	Data warehouses
users	thousands	hundreds
workload	predefined transactions	<i>ad hoc</i> analysis queries
access	to hundreds of records, read and write	to millions of records, mostly read-only
goal	application-dependent	decision support
data	elementary, numeric and alphanumeric	aggregated, mostly numeric
data integration	application-based	subject-based
quality	in terms of integrity	in terms of consistency
temporal span	current data	current and historical data
update	continuous	periodic
model	normalized	multidimensional
optimization	for OLTP accesses on a fraction of database	for OLAP accesses on a large part of database

Two-layer architectures



DATA MART:

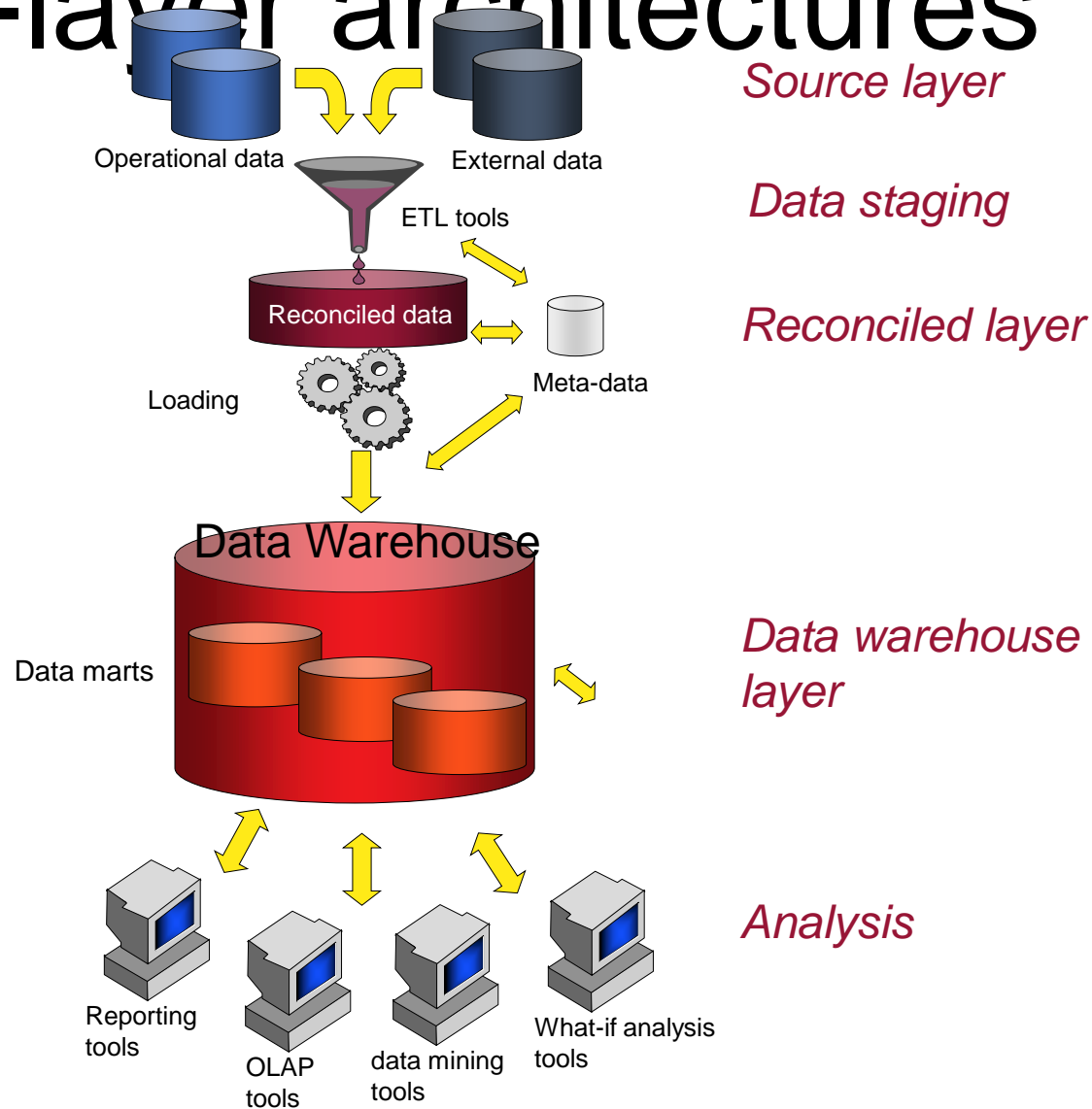
A subset or an aggregation of the data stored to a primary data warehouse. It includes a set of information pieces relevant to a specific business area, corporate department, or category of users.

Two-layer architectures

Pros:

- At the level of the warehouse a quality information is continuously available even when, for technical or organizational issues, access to the sources is temporarily denied
- The analytical queries performed on the DW does not interfere with the handling of transactions at the operational level, whose reliability is essential to business operations
- The logical organization of the DW is based on the multidimensional model, while the sources typically offer relational models
- There is temporal and granularity discrepancy between OLTP systems, handling current at the maximum level of detail, and OLAP systems that operate on historical and summarized data
- At the DW level you can use specific techniques to optimize applications for performance analysis and reporting

Three-layer architectures



RECONCILED DATA:

operational data obtained after integrating and cleansing source data. As a result, those data are integrated, consistent, appropriate, current, and detailed

Three-layer architectures

The main advantage of the reconciled data level is that it creates a common data model and reference for the whole company, while introducing a clear distinction between issues related to the extraction and integration of data from sources and those inherent the DW feeding

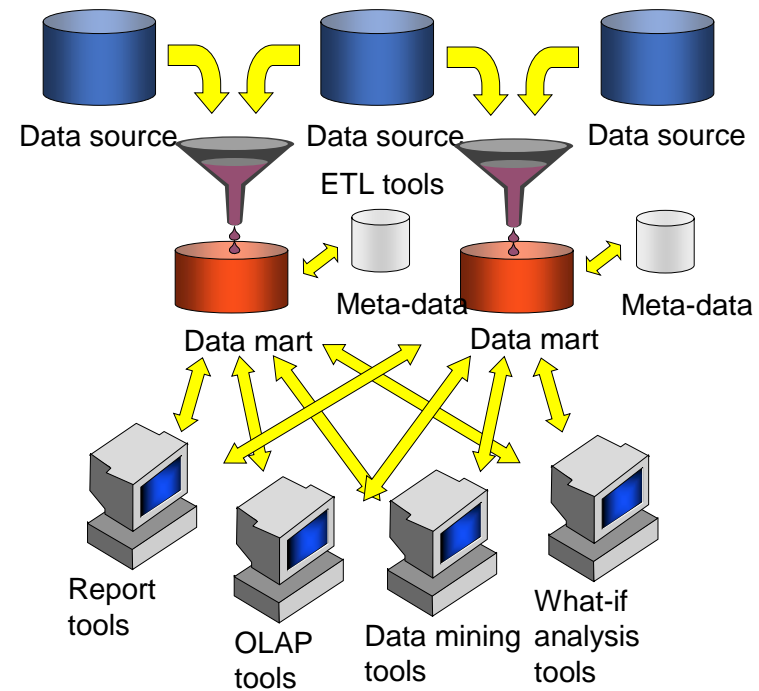
On the other hand, reconciled data introduce additional redundancy compared with source operational data

Architectures: another classification

- Data mart independent
- Data mart bus
- Hub-and-spoke
- Centralized data warehouse
- Federation

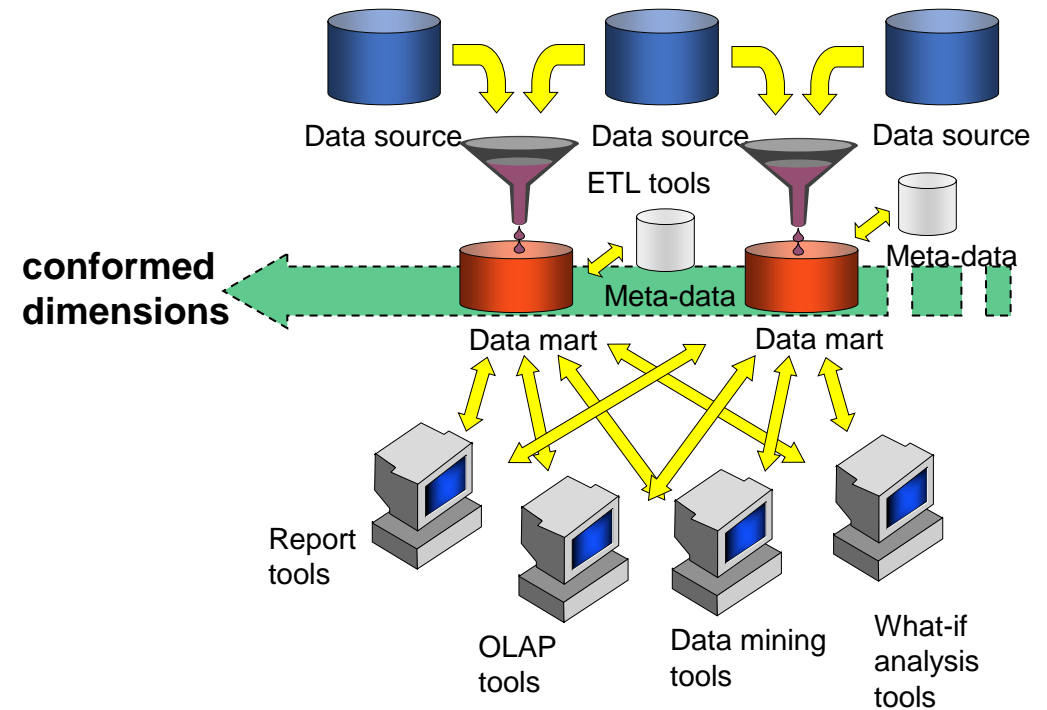
Data mart independent

First approach to data warehousing
Inconsistency issues



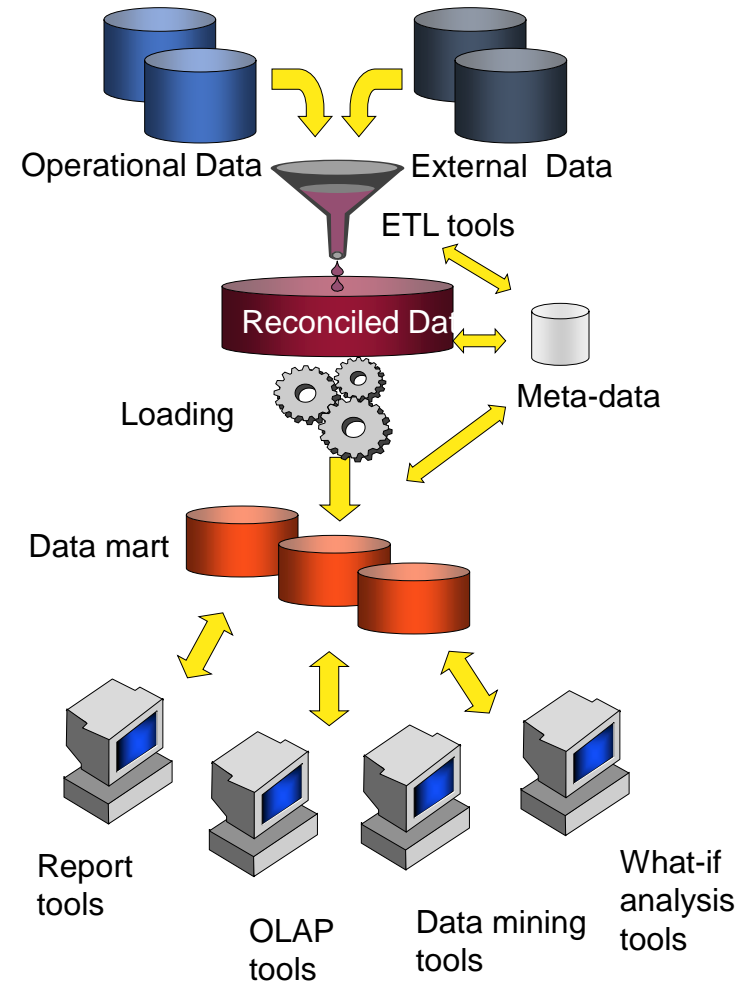
Data mart bus

Approach suggested by [Kimball](#)
Logical level integration
“Enterprise view”



Hub-and-spoke

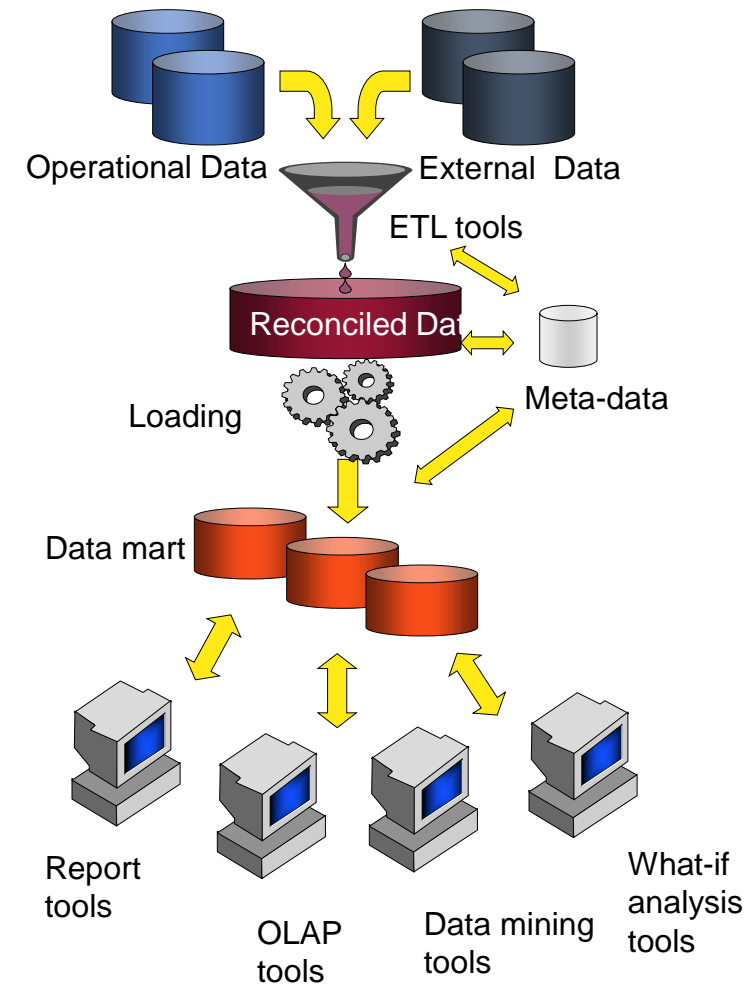
One of the most used architectures in medium to large environments



Centralized Data warehouse

Approach suggested by [Inmon](#)

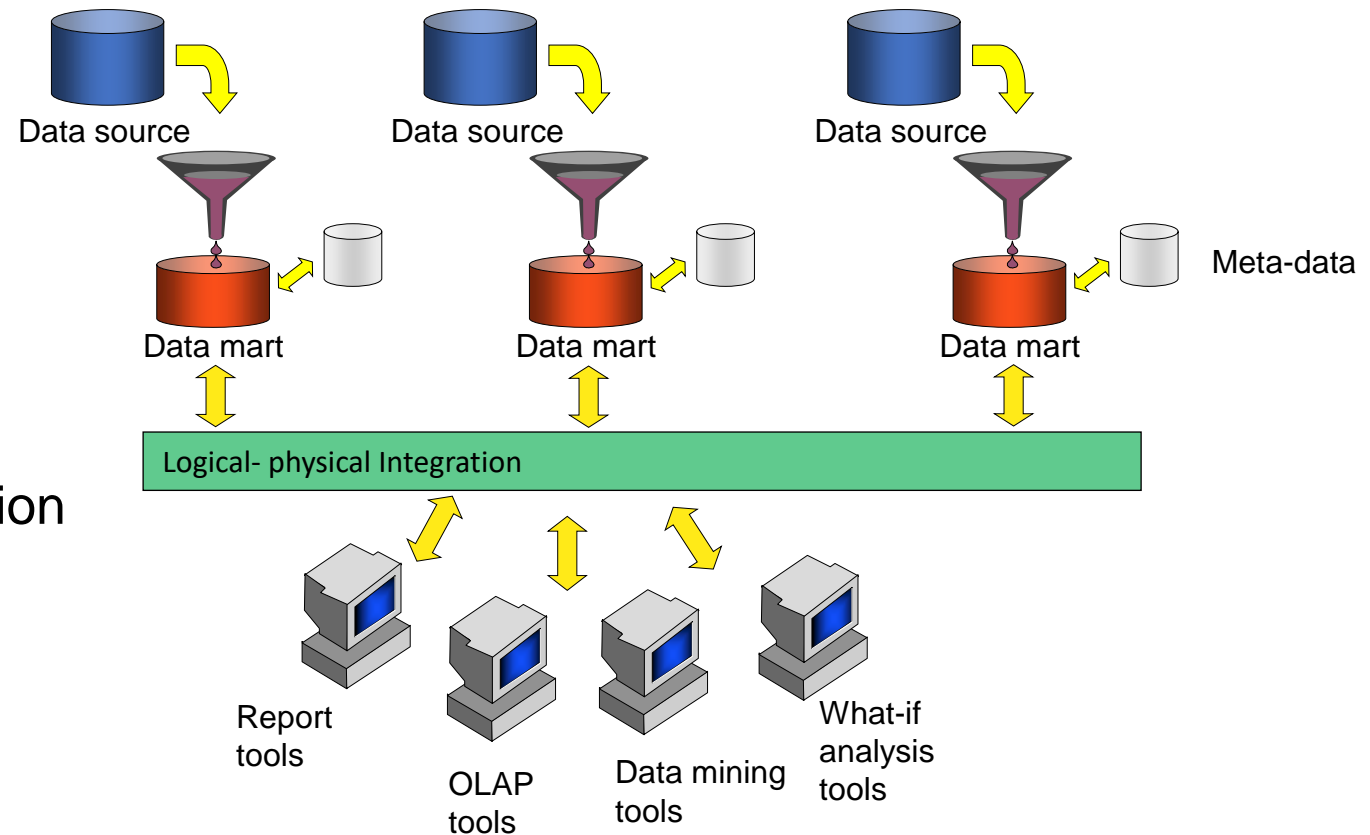
Logical, more than physical, implementation
of a hub-and-spoke architecture



Federation

Good for highly dynamic environments
(mergers and acquisitions)

Problem of effective and efficient integration



Choosing between architectures

Information interdependence among organizational units in company

- encourages the adoption of enterprise-wide architectures

Urgency of the data warehousing project

- encourages the adoption of “fast” architectures

Constraints on economic and human resources

Role of the project within the business strategy

- independent data marts vs. hub-and-spoke

Compatibility with existing platforms

Skills of the IT staff

Organizational position of the sponsor of the project

- enterprise architectures vs. departmental architectures

Towards the multidimensional model

- “What business were registered last year for each region and each product category?”
- “What is the relationship between performance of shares of PC manufacturers and quarterly profits along the last 5 years?”
- “What are the types of orders that will maximize revenues?”
- “Which of two new therapies is more effective for decreasing the average length of hospitalizations?”
- “What is the relationship between the profits made with shipments of less than 10 items and those made with shipments of more than 10 items?”

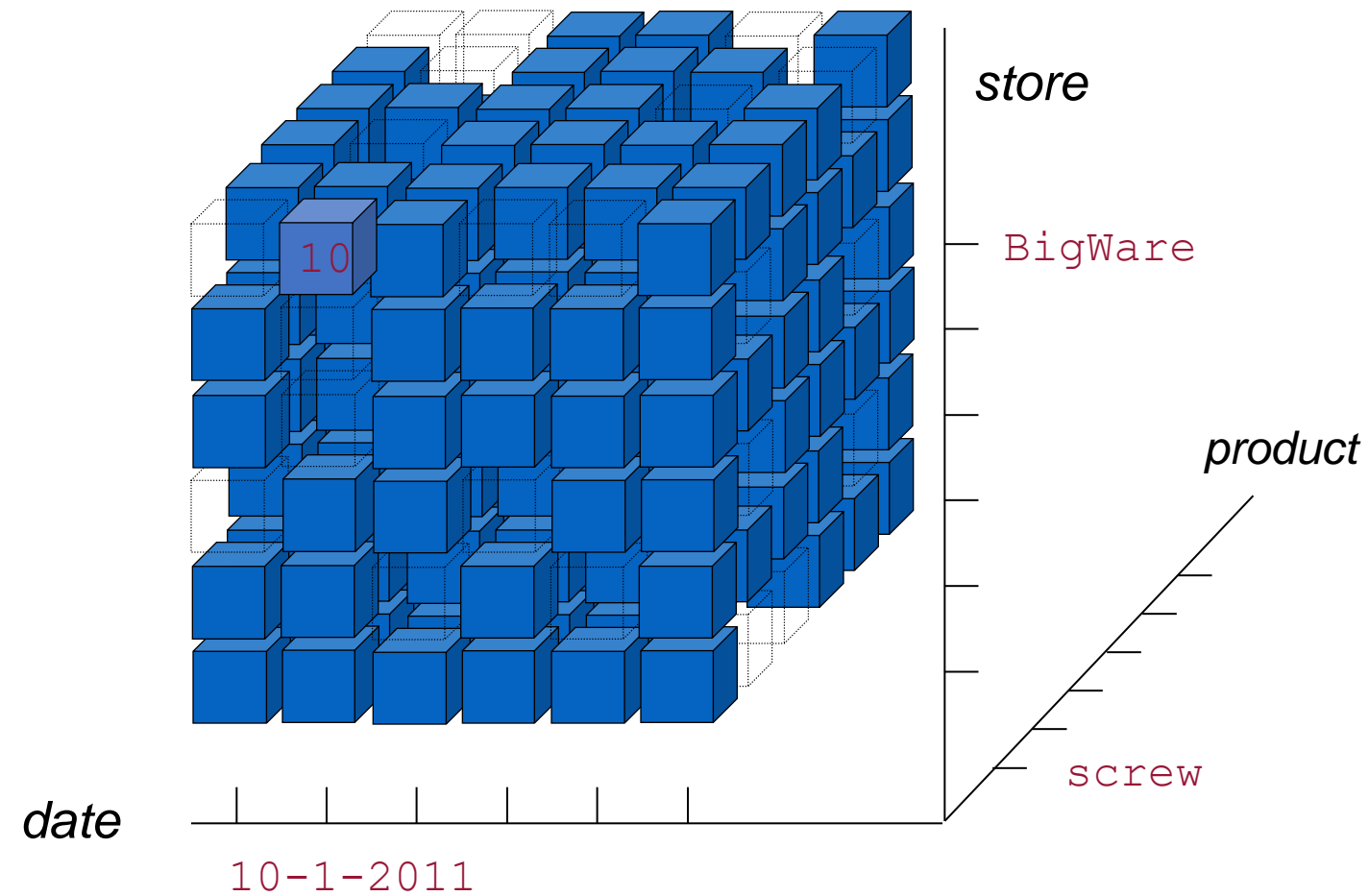
The multidimensional model

It is the key for representing and querying information in a DW

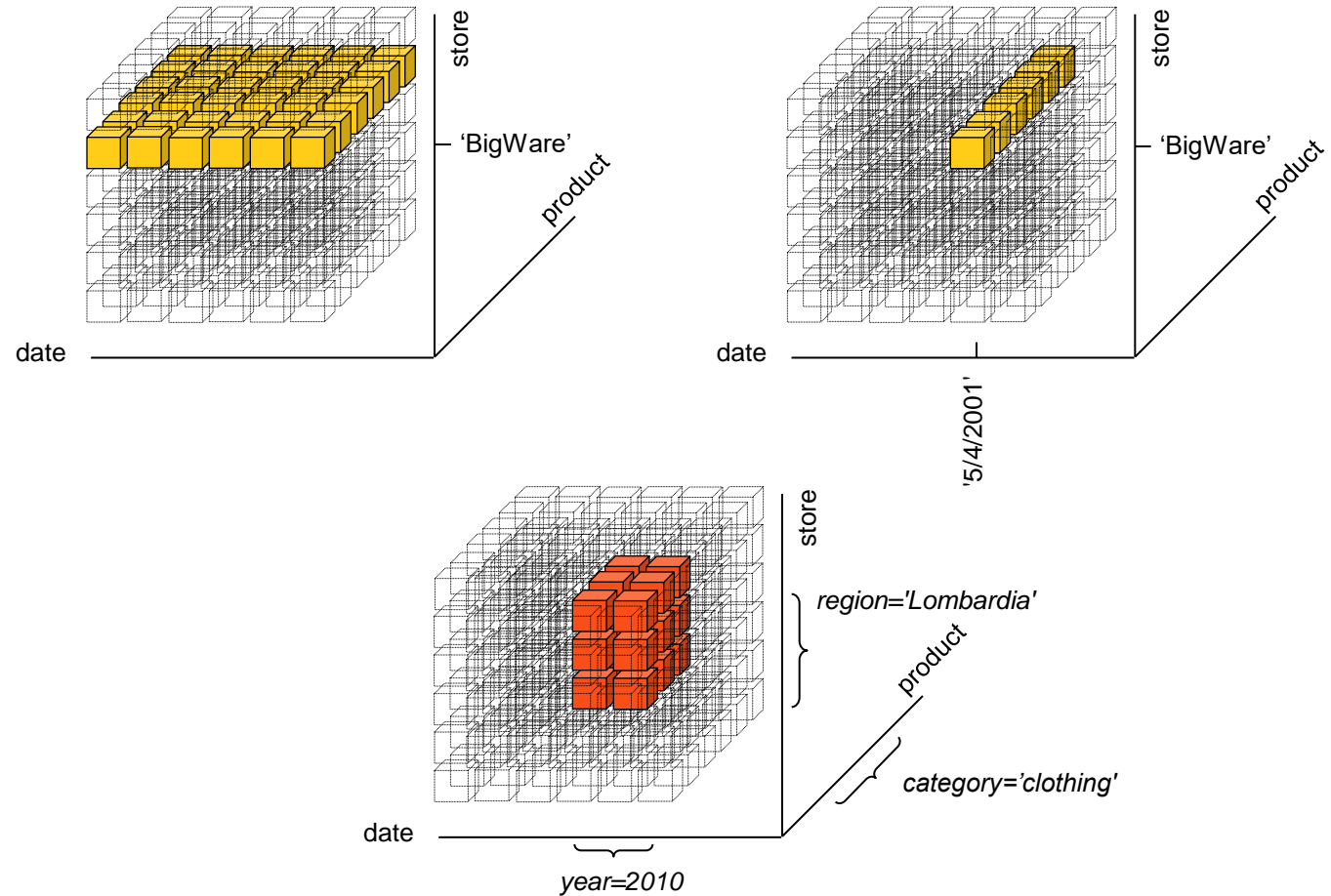
Facts of interest are represented in *cubes* where:

- each cell stores numerical *measures* that quantify the fact from different points of view;
- each axis is a *dimension* for analyzing measure values;
- each dimension can be the root of a *hierarchy* of attributes used to aggregated measure values

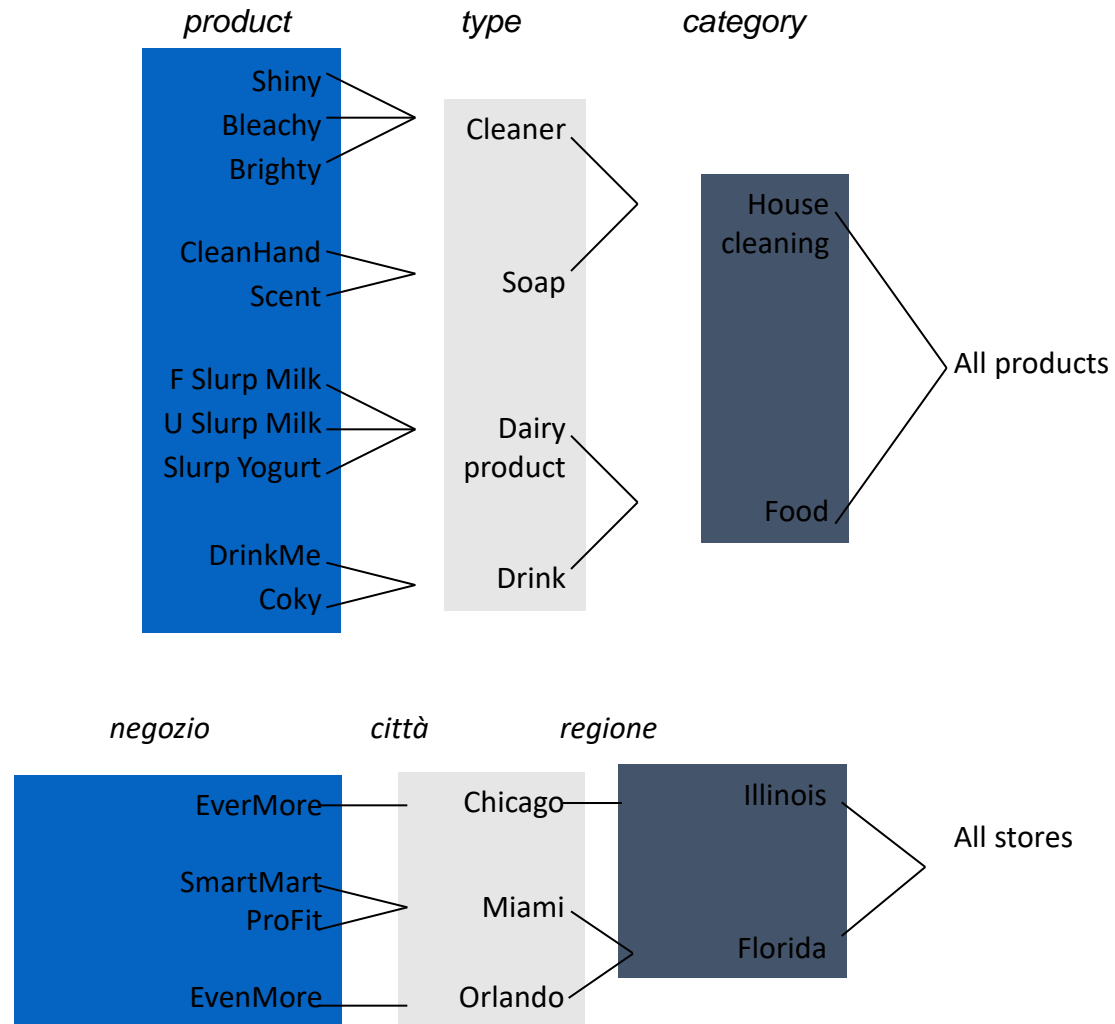
The Sales cube



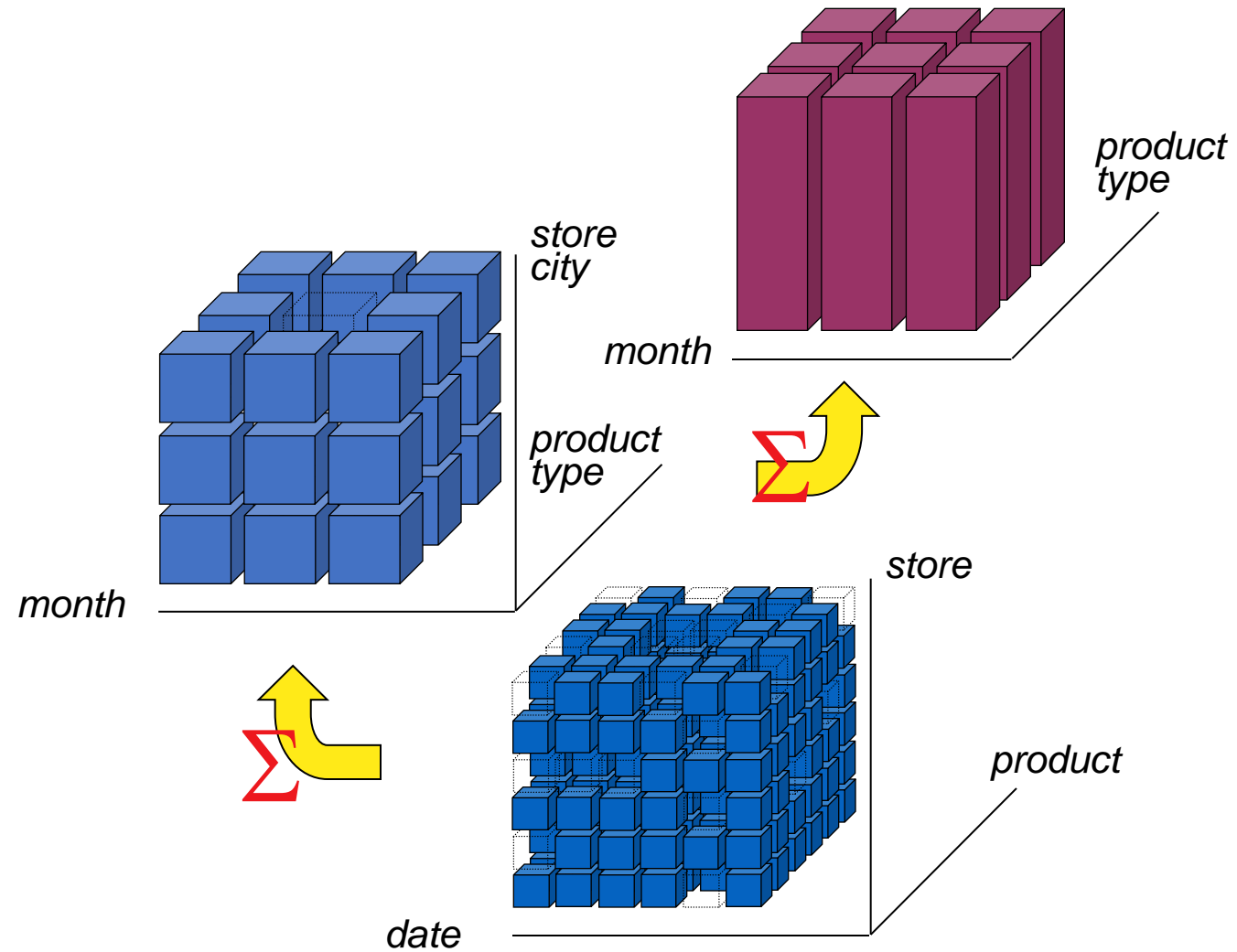
Slicing and dicing



Hierarchies

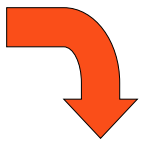


Aggregation

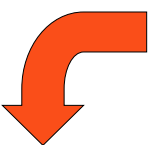


Aggregation

	BigWare1	BigWare2	NotOnlyJelly
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
.....



	BigWare1	BigWare2	NotOnlyJelly
January 2000	200	180	150
February 2000	180	150	120
March 2000	220	180	160
.....
January 2001	350	220	200
February 2001	300	200	250
March 2001	310	180	300
.....
January 2002	380	200	220
February 2002	310	200	250
March 2002	300	160	280
.....



	BigWare1	BigWare2	NotOnlyJelly
2000	2400	2000	1600
2001	3200	2300	3000
2002	3400	2200	3200



	BigWare1	BigWare22	NotOnlyJelly
Totale:	9000	6500	7800

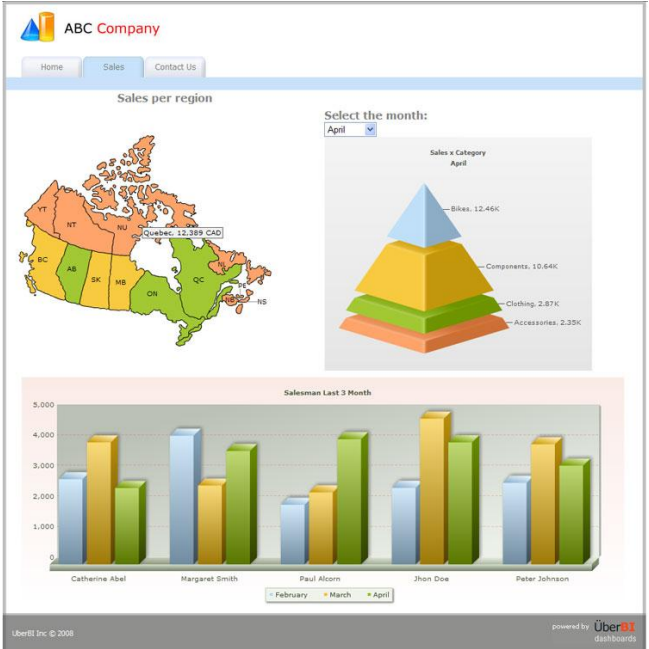
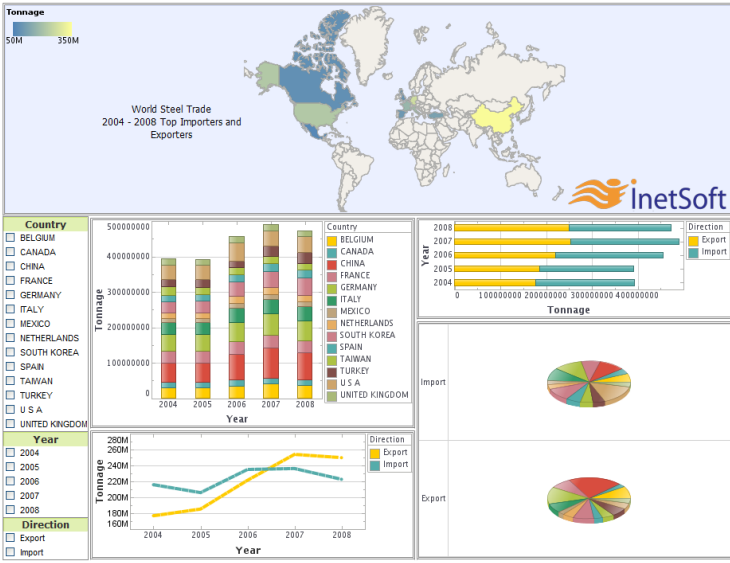
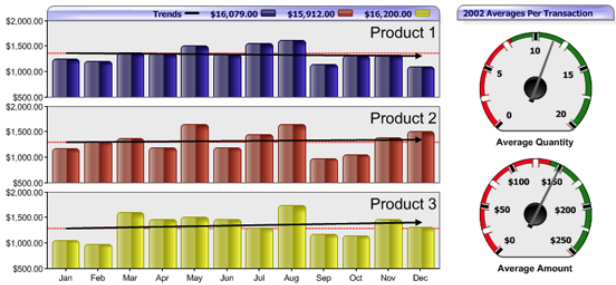
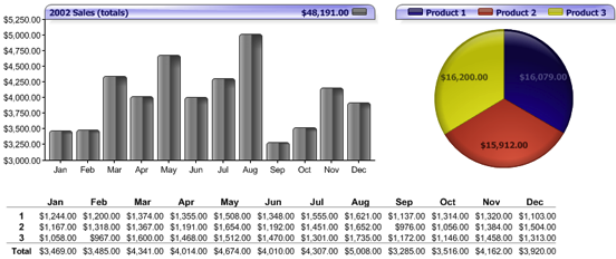
Data analysis techniques

Once data have been cleaned, integrated, and transformed, users must be enabled to take maximum advantage from the resulting information assets

There are two different approaches for querying data warehouses, supported by different types of tools:

- **Reporting**: no ICT skills required
- **OLAP**: users must be able to reason according to the multidimensional paradigm, and they must be acquainted with the visual interface of the tool

Reporting



OLAP

OLAP is **the main way** to exploit information in a data warehouse

It gives end-users, whose analysis needs are not easy to define beforehand, the opportunity to **analyze and explore data interactively** on the basis of the multidimensional model

While users of reporting tools play a passive role, OLAP users are able to **actively** carry out a complex analysis session, where each step is the result of the outcome of preceding steps

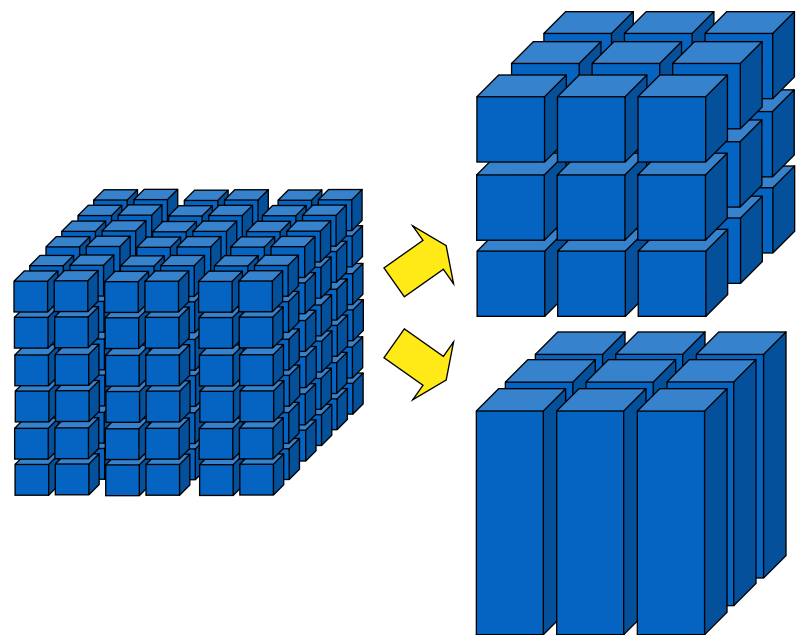
OLAP session

An OLAP session consists of a **navigation path** that corresponds to an analysis process for facts according to different viewpoints and at different detail levels. This path is turned into a sequence of queries, which are not issued directly, but differentially expressed with reference to the previous query

Every step of an analysis session is characterized by an **OLAP operator** that turns the latest query into a new one

The **results of queries are multidimensional**; OLAP tools typically use tables to display data, with multiple headers, colors, and other features to highlight data dimensions

OLAP operators



roll-up

OLAP operators

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



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

OLAP operators

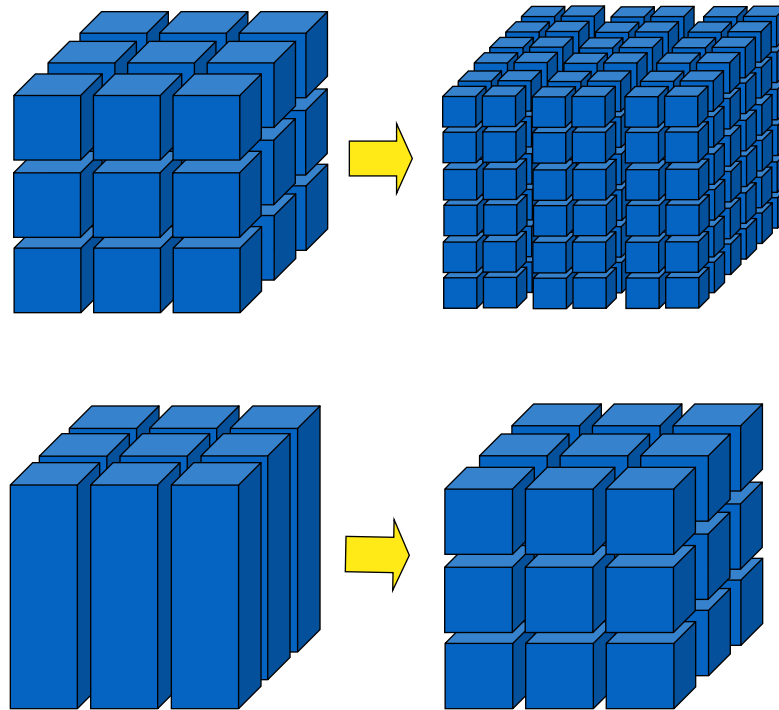
Category	Year	Metrics Customer Region	Dollar Sales									
			North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germa
Electronics	1997		\$ 138	\$ 1.774	\$ 384	\$ 138	\$ 2.346	\$ 2.554	\$ 2.184	\$ 566	\$ 199	\$
	1998		\$ 1.184	\$ 4.529	\$ 1.892	\$ 7.232	\$ 651	\$ 9.488	\$ 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	\$

roll-up



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

OLAP operators



drill-down

OLAP operators

	Metrics	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	\$ 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



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

OLAP operators

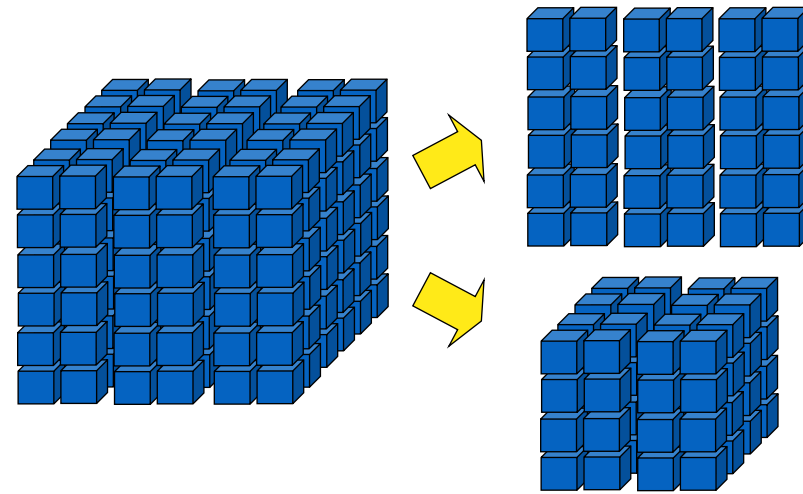
	Metrics	Dollar Sales	
	Year	1997	1998
Category			
Electronics		\$ 10.616	\$ 29.299
Food		\$ 5.300	\$ 5.638
Gifts		\$ 16.315	\$ 20.047
Health & Beauty		\$ 6.042	\$ 5.665
Household		\$ 38.383	\$ 50.391
Kid's Korner		\$ 2.559	\$ 2.943
Travel		\$ 4.497	\$ 4.792



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

drill-down

OLAP operators



slice-and-dice

OLAP operators

		Metrics Customer Region	Dollar Sales									
			North-East	Mid-Atlantic	South-East	Central	South	North-West	South-West	England	France	Germa
Category	Year											
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	\$

slice-and-dice



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	

OLAP operators

	Metrics	Dollar Sales											
	Customer City	Afton	Akron	Albon	Alcameda	Alka	Allagash	Alta	Altoola	Amestra	Amsterdam	Andersonville	Annap
Subcategory													
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			\$:

slice-and-dice



Filter Details:						
Category = Electronics						
AND						
Dollar Sales > 80						
AND						
Customer Region = North-West						
AND						
Year = 1997						
	Metrics	Dollar Sales				
	Customer City	Alta	Armstrong	Avery Heights	Lane	Mt. Everest
Subcategory						
Audio			\$ 98		\$ 123	\$ 85
Comfort				\$ 118		\$ 1.495
Gadgets		\$ 199				\$ 199

OLAP tools

OLAP tools were born at the end of the 1990s and today continue to renew and include new features



Live time!

Methodological framework

Why?

Building a DW is a very complex task, which requires an **accurate planning** aimed at devising satisfactory answers to organizational and architectural questions

A large number of organizations lack the experience and skills required to meet the **challenges** involved in DW projects

The reports of DW project failures state that a major cause lies in the absence of a global view of the design process: in other terms, in **the absence of a design methodology**

Methodologies are created by closely studying similar experiences and **minimizing the risks for failure** by basing new approaches on a constructive analysis of the mistakes made previously

Risk factors

- Risks related to project management
- Risks associated with technology
- Risks related to data and design
- Risks related to the organization

The risk of getting an unsatisfactory result in data warehousing projects is particularly high because of high user expectations

In business contemporary culture there is a widespread belief that attaches to data warehousing the role of *panacea* (cure-all)

Actually a large part of the responsibility of the project's success falls on the quality of the data sources and on foresight, dynamism, and availability of company staff

Top-down approach

Analyze global business needs, plan how to develop a data warehouse, design it, and implement it as a whole

- 👍 This procedure is promising: it is based on a global picture of the goal to achieve, and in principle it ensures consistent, well integrated data warehouses
- 👎 High-cost estimates with long-term implementations discourage company managers from embarking on these kind of projects
- 👎 Analyzing and integrating all relevant sources at the same time is a very difficult task, even because it is not very likely that they are all available and stable at the same time
- 👎 It is extremely difficult to forecast the specific needs of every department involved in a project, which can result in the analysis process coming to a standstill
- 👎 Since no working system is going to be delivered in the short term, users cannot check for this project to be useful, so they lose trust and interest in it

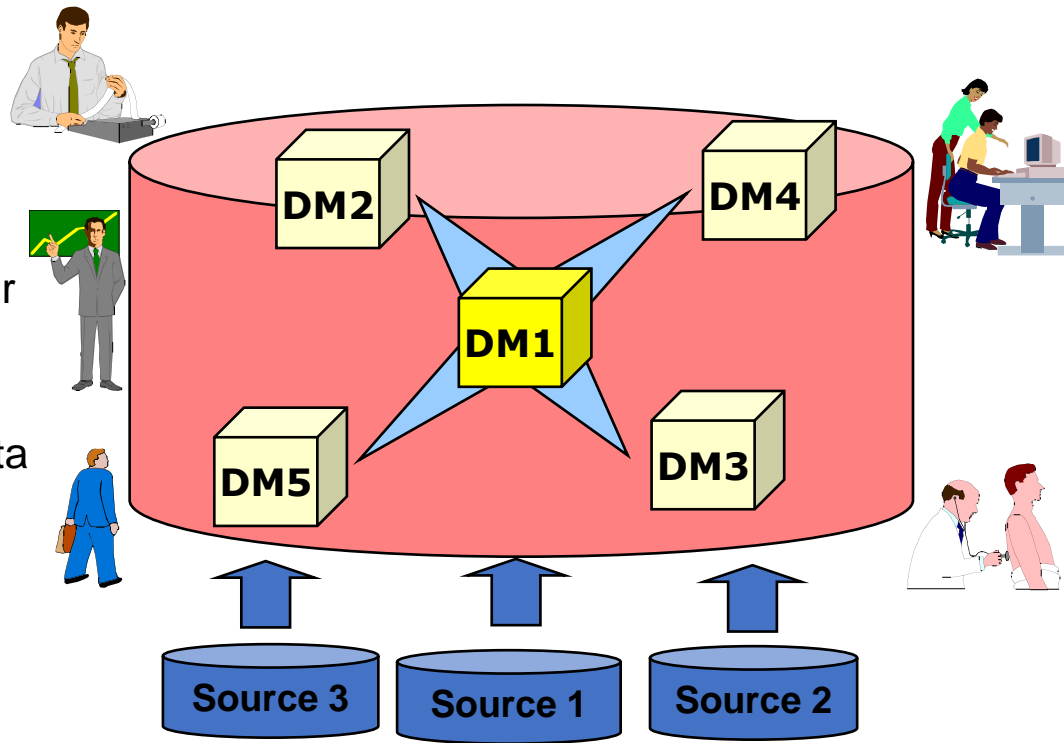
Bottom-up approach

DWs are incrementally built and several data marts are iteratively created. Each data mart is based on a set of facts that are linked to a specific department and that can be interesting for a user group

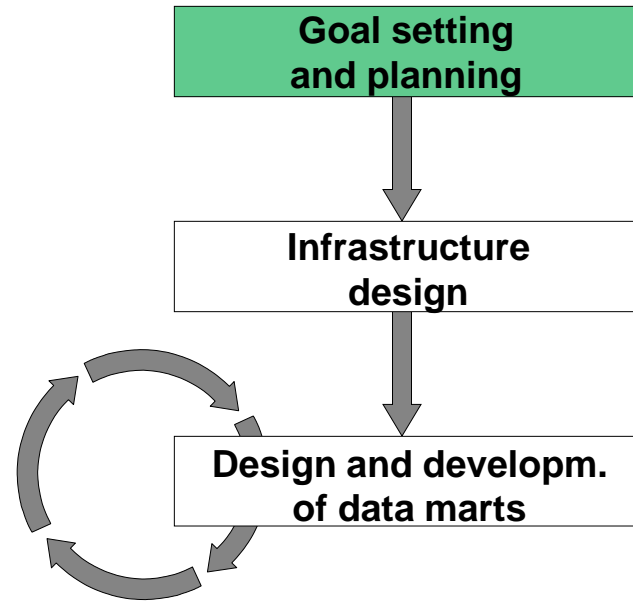
- 👍 Leads to concrete results in a short time
- 👍 Does not require huge investments
- 👍 Enables designers to investigate one area at a time
- 👍 Gives managers a quick feedback about the actual benefits of the system being built
- 👍 Keeps the interest for the project constantly high
- 👎 May determine a partial vision of the business domain

The first data mart to be prototyped...

- is the one playing the most strategic role for the enterprise
- should be a backbone for the whole DW
- should lean on available and consistent data sources

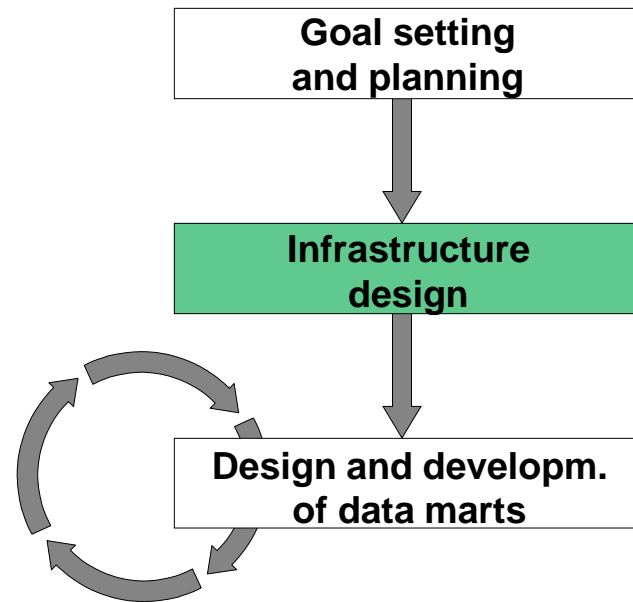


The life-cycle



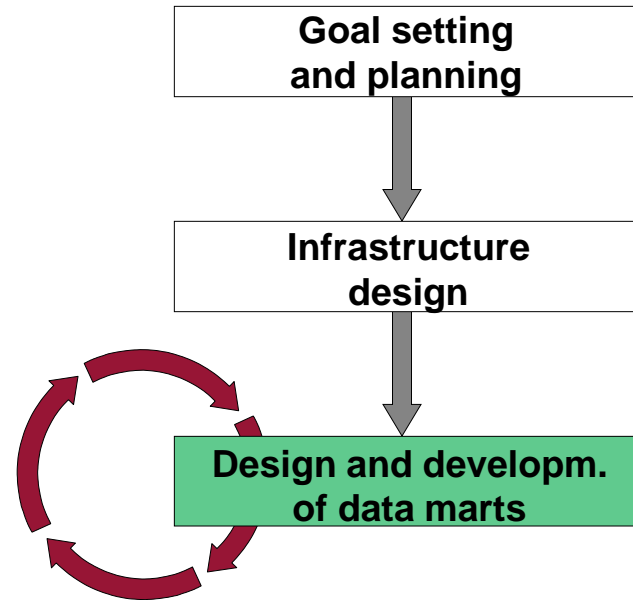
- set system goals, borders, and size
- select an approach for design and implementation
- estimate costs and benefits
- analyze risks and expectations
- examine the skills of the working team

The life-cycle



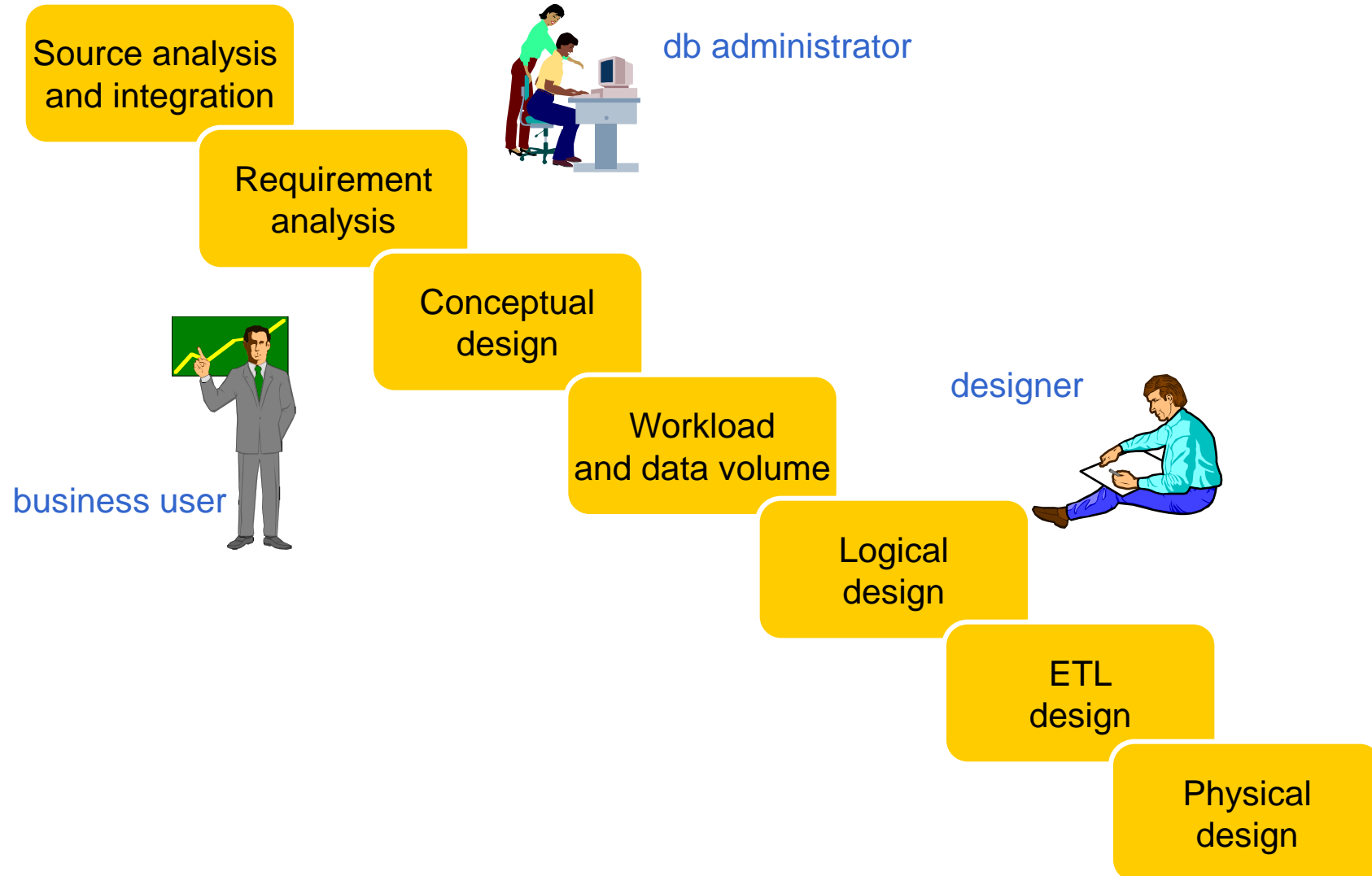
- analyze and compare the possible architectural solutions
- assess the available technologies and tools
- create a preliminary plan of the whole system

The life-cycle



Every iteration causes a new data mart and new applications to be created and progressively added to the DW system

Data mart design phases



Methodological scenarios

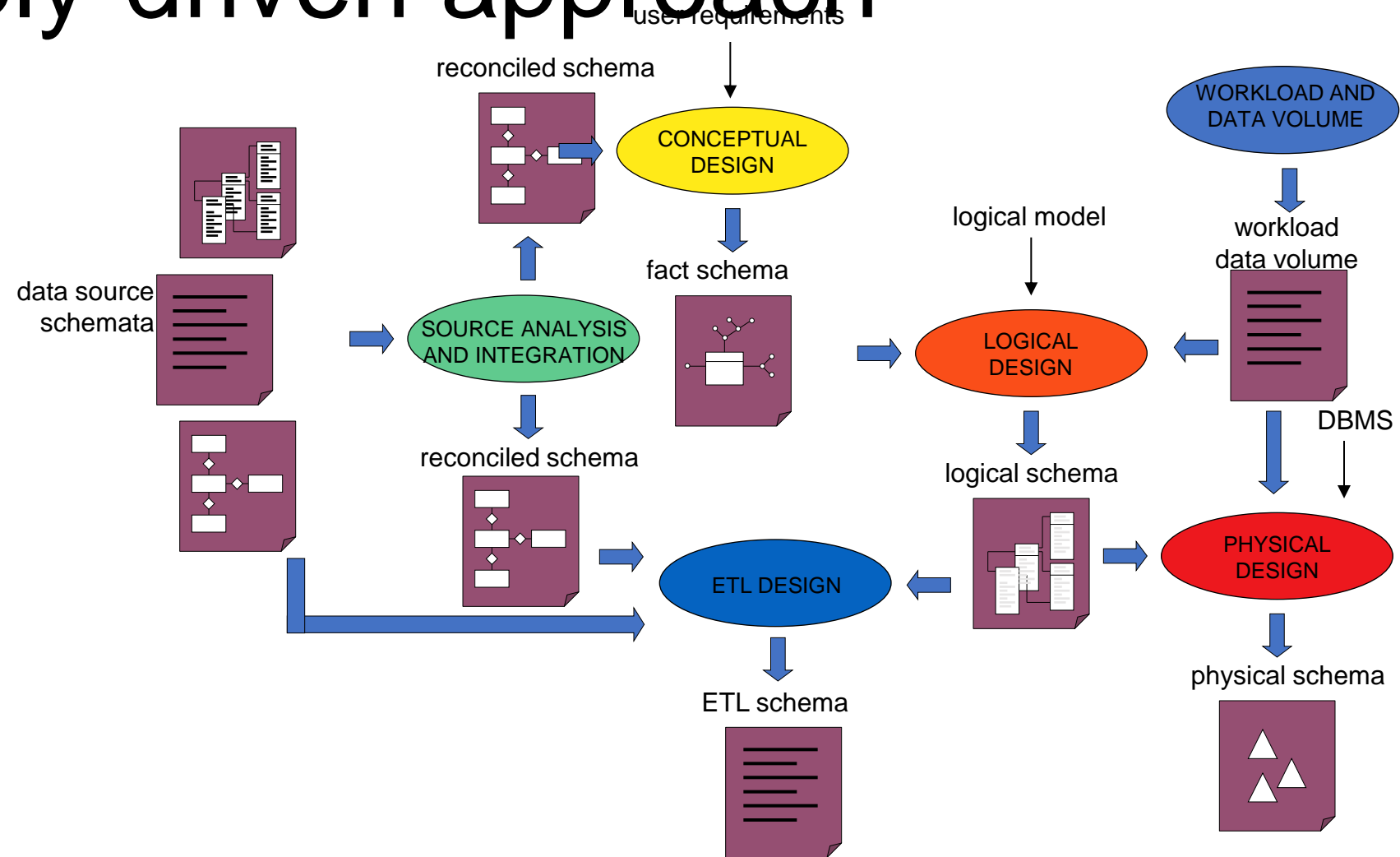
Supply-driven approach

- they design data marts on the basis of a close operational data source analysis
- user requirements show designers which groups of data, relevant for decision-making processes, should be selected and how to define data group structures on the basis of the multidimensional model

Demand-driven approach

- they begin with the definition of information requirements of data mart users
- the problem of how to map those requirements into existing data sources is addressed at a later stage, when ETL procedures are implemented

Supply-driven approach



Supply-driven approach

Pros

- an initial conceptual schema for data marts can be **automatically derived** from the reconciled layer—that is, it strictly depends on data source structures
- ETL design is **extremely streamlined** because every single information piece stored in a data mart is directly associated with one or more source attributes
- the resulting data marts are quite **stable in time**, because they are rooted in source schemata—that change less frequently than the requirements expressed by end users

Cons

- user requirements play a minor role when it comes to specifying information contents to carry out an analysis
- designers have a limited support when facts, dimensions, and measures need to be determined

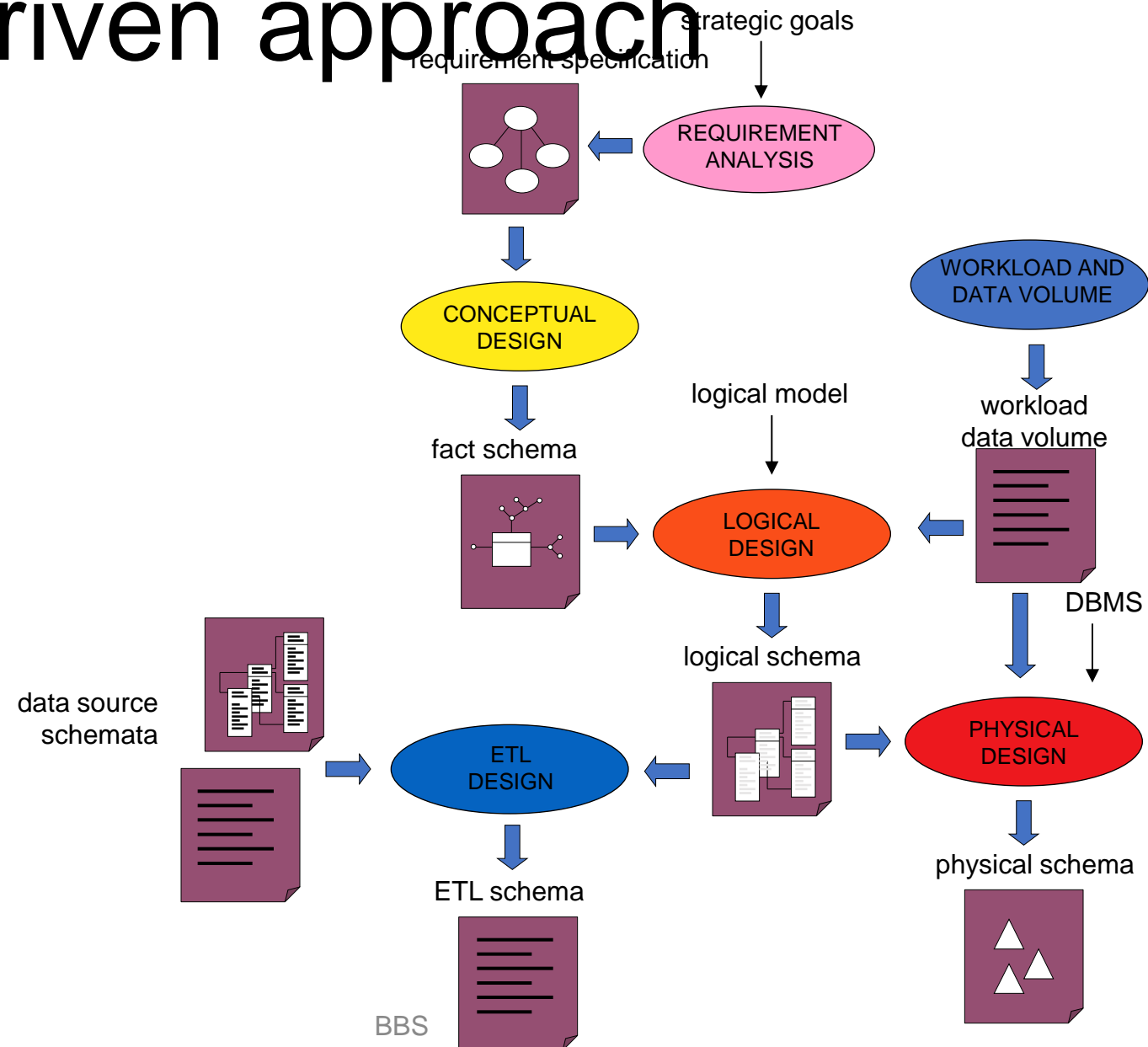
Supply-driven approach

Applicability

- It can be applied when:
 - An in-depth knowledge of data sources to populate data marts is available or it can be achieved at a reasonable price and in the short term
 - Data source schemata show a good level of normalization
 - Data source schemata are not too complex
- When your architecture includes a reconciled layer, you can amply meet those requirements: normalization and in-depth knowledge are obtained at the end of the reconciliation process
- If your data sources are reduced to one single, small, well-designed database, you can achieve the same result after accurately completing the inspection phase

Our experience in design shows that the data-driven approach, if applicable, is better than other approaches, because it gives you the opportunity to reach your project goals within a very short time

Demand-driven approach



Demand-driven approach

Pros

- users' wishes play a **leading role**

Cons

- designers are required to have strong leadership and meeting facilitation qualities to properly grab and integrate the different points of view
- designers make great efforts in the data-staging design phase
- facts, measures, and hierarchies are drawn directly from the specifications provided by users, and only at a later stage can designers check for the information required to be actually available in source databases
- this may undermine customers' confidence in designers and in the advantage gained by data marts on the whole

Demand-driven approach

Applicability

- This is **your only alternative** if you cannot conduct any preliminary, detailed source analysis (for example, when an ERP system is used to feed your data mart) or if sources are represented by legacy systems with such complexity that it is not recommended that you explore and normalize them, and as a result you do not think it appropriate to create the reconciled layer

Demand-driven approaches are typically more time-expensive than data-driven approaches, because users often do not have a clear and shared understanding of business goals and processes

A more agile approach...

Agile methodologies can only partially apply to DWs

- Focus on incremental delivery of functionalities that are relevant for the users...
 - ... many DW modules are hardly perceived as useful by the users (e.g., the ETL procedures)
- Very segmented projects based on user stories: functional requirements that can be implemented in a few days...
 - ... A design approach based on detailed user requirements (typically single report) might not bring out the real multi-dimensional data structure

Nevertheless, many of the underlying principles of agile methodologies can be reused

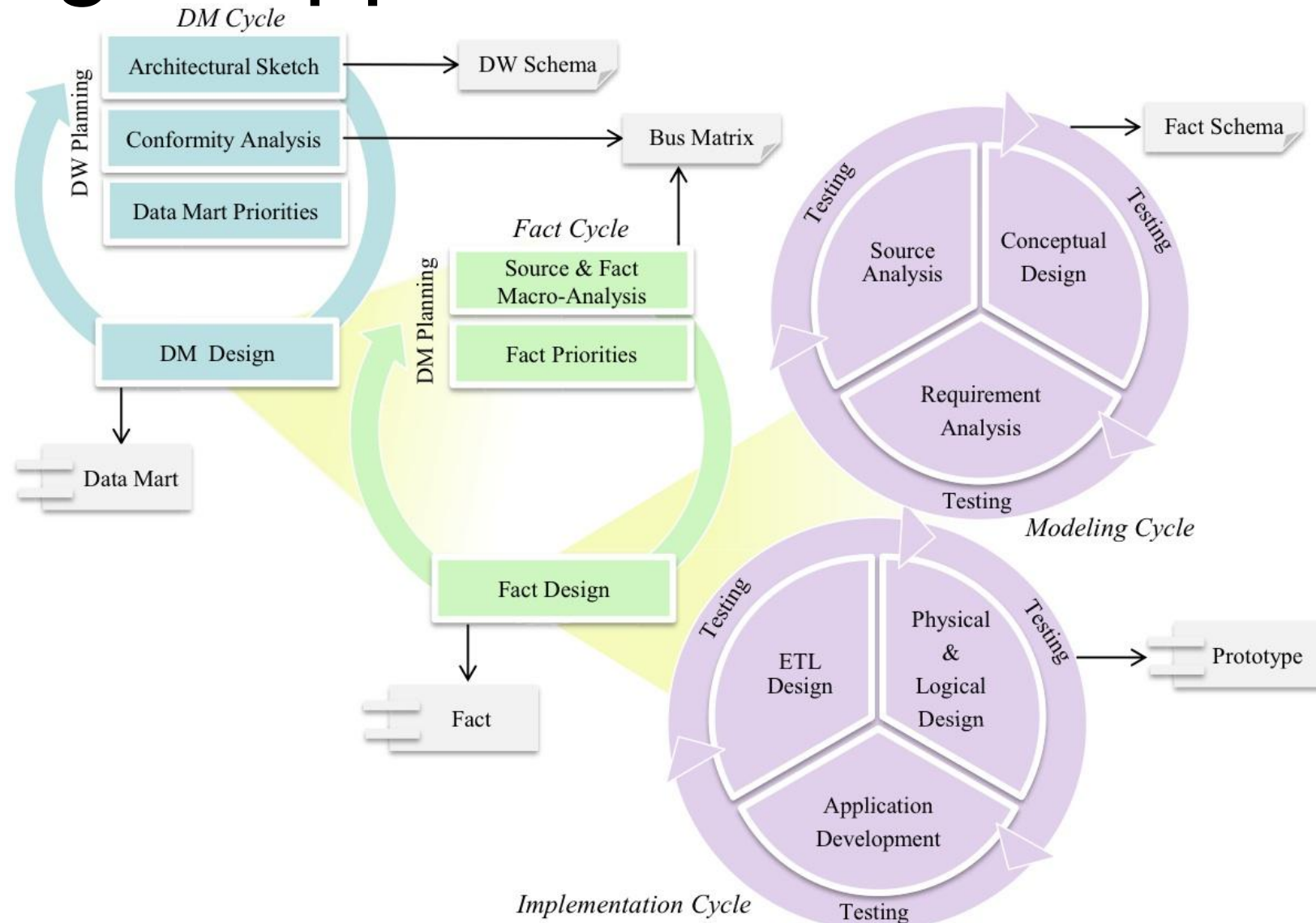
- Incrementality and risk-based iteration
- Prototyping
- Strong user involvement
- Lightweight and formal documentation
- Reuse of components
- Automatic generation of diagrams

A more agile approach...

The methodology that emerges from the use of agile principles in DW real projects includes:

- Iteration cycles based on facts
- Creation of a conceptual model before the implementation phase
- User participation in testing activities
- Willingness to review the project's priorities in terms of facts and data marts

A more agile approach...



Conceptual design

Which formalism?

While it is now universally recognized that a data mart is based on a multidimensional view of data, there is still **no agreement** on how to implement its conceptual design

Use of the **Entity-Relationship model** is quite widespread throughout companies as a conceptual tool for standard documentation and design of relational databases, but *it cannot be used to model DWs*

In some cases, designers base their data marts design on the logical level—that is, they directly define **star schemata** that are the standard ROLAP implementation of the multidimensional model. But a star schema is nothing but a relational schema; *it contains only the definition of a set of relations and integrity constraints!*

The Dimensional Fact Model

The DFM is a graphical conceptual model for data mart design, devised to:

1. lend effective support to conceptual design
2. create an environment in which user queries may be formulated intuitively
3. make communication possible between designers and end users with the goal of formalizing requirement specifications
4. build a stable platform for logical design (*independently of the target logical model*)
5. provide clear and expressive design documentation

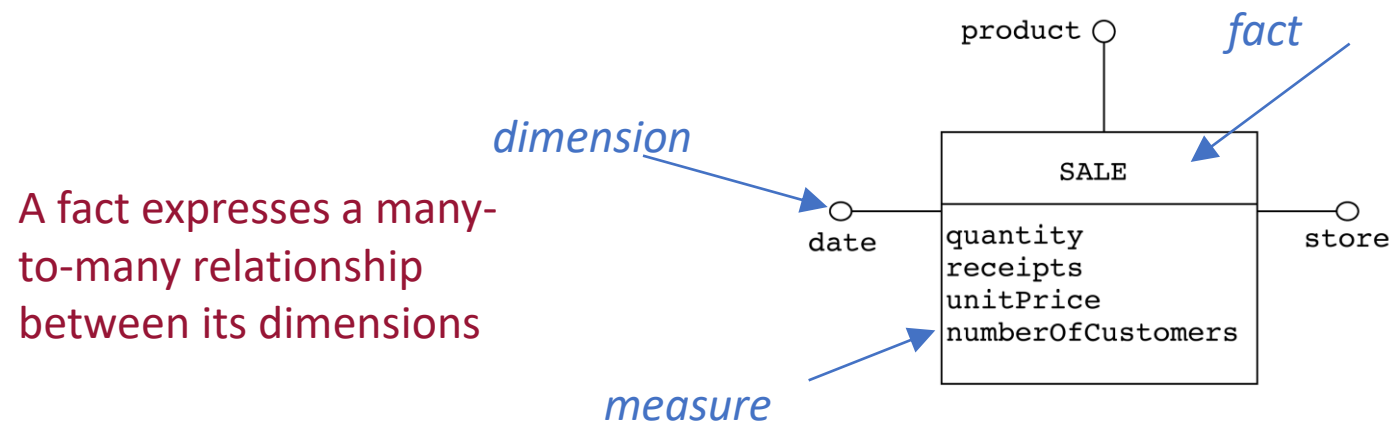
The conceptual representation generated by the DFM consists of a set of *fact schemata* that basically model facts, measures, dimensions, and hierarchies

DFM: basic concepts

A **fact** is a concept relevant to decision-making processes. It typically models a set of events taking place within a company (e.g., sales, shipments, purchases, ...). It is essential that a fact have dynamic properties or evolve in some way over time

A **measure** is a numerical property of a fact and describes a quantitative fact aspect that is relevant to analysis (e.g., every sale is quantified by its receipts)

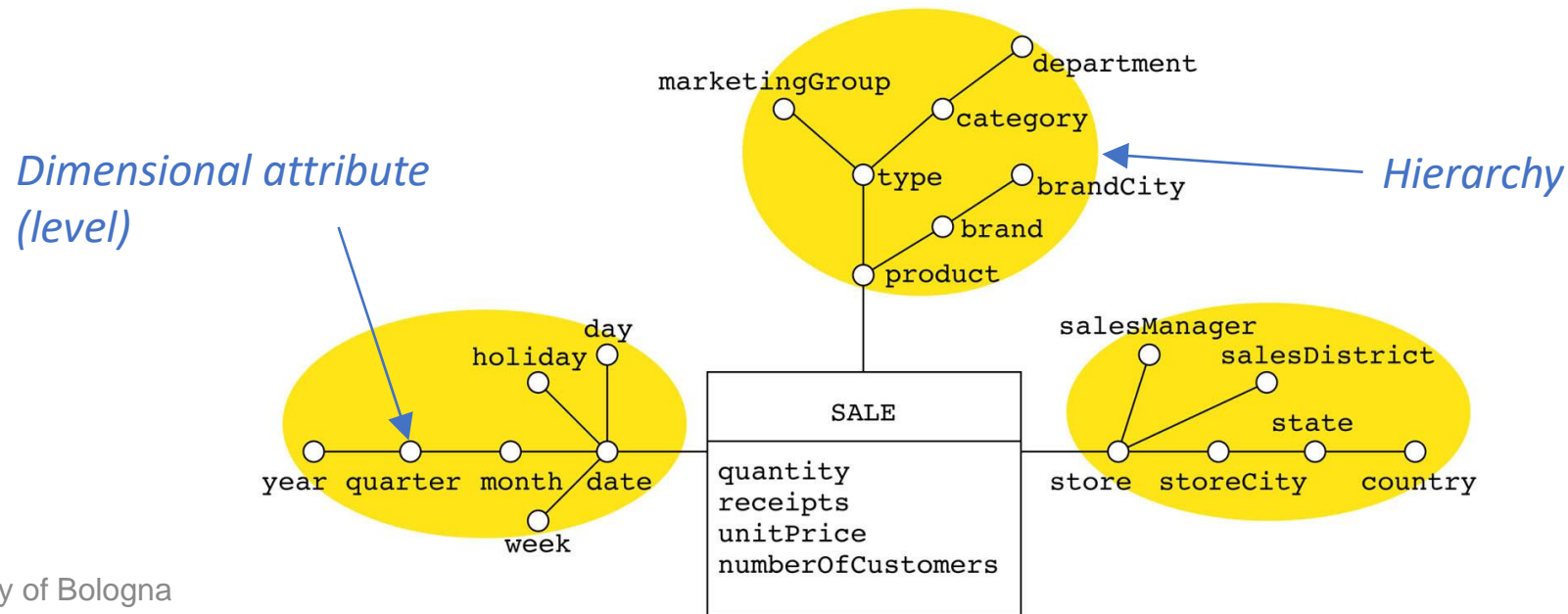
A **dimension** is a fact property with a finite domain and describes an analysis coordinate of the fact. Typical dimensions for the sales fact are products, stores, and dates



DFM: basic concepts

The general term **dimensional attributes** stands for the dimensions and other possible attributes, always with discrete values, that describe them (e.g., a product is described by its type, by the category to which it belongs, by its brand, and by the department in which it is sold)

A **hierarchy** is a directed tree whose nodes are dimensional attributes and whose arcs model many-to-one associations between dimensional attribute pairs. It includes a dimension, positioned at the tree's root, and all of the dimensional attributes that describe it



DFM vs. ERM

