Data Warehouse

NETTAB workshop 2011

Carlo Combi
Elena Gaspari
Alberto Sabaini

Department of Computer Science, University of Verona, Verona, Italy

Outline

Outline

Introduction

Definition

Information = value-increasing asset, needed to effectively plan and control decision-based activities, as diagnosis, therapy planning, monitoring, health care management.

Unfortunately data $\neq$ information.

Having a huge amount of data makes it difficult to extract useful information.

Data Warehousing process was born to handle this huge amount of data that increased in this last decade.

Mixing together analytical and transactional queries leads to inevitable delays.

Basic Idea: to separate *On-Line Analytical Processing* (OLAP) from *On-Line Transactional Processing* (OLTP), building a new collector of information that integrates data from different sources i.e., the Data Warehouse.

Outline

Data Warehousing

Definition

Decision Support System: set of techniques and software tools to extract information from a set of data stored in different sources.

Among the Decision Support Systems, **Data Warehouse Systems** are those that are more established in the industrial world and could be suitably used also for biomedical data.

Definition

Data Warehousing: a collection of methods, technologies and tools to assist the *knowledge worker* (clinician, manager, nurse, epidemiologist, technician) to perform data analysis aimed at improving decision making and information assets.

Complaints

- We have a huge amount of data but we can not access it!
- Why people doing the same role are showing significantly different results?
- We want to select, combine and manipulate data in every possible way!
- Show me only what is important!
- Everyone knows that some data are not correct!

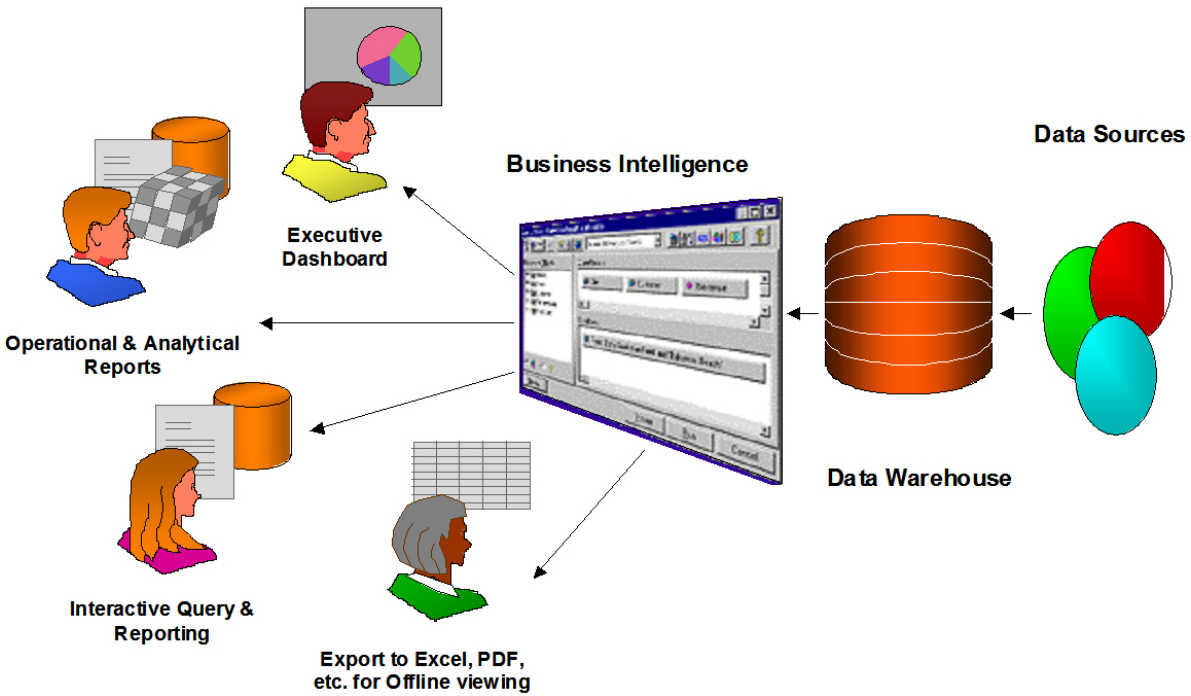
Characteristics of the Warehousing process

- Accessibility to users with limited knowledge of computing and data structures.
- Data integration based on a standard model.
- Flexible query to take full advantage of the wealth of information.
- Synthesis to allow targeted and effective analysis.
- Multidimensional representation to provide an intuitive view of information.
- Correctness and completeness of integrated data.

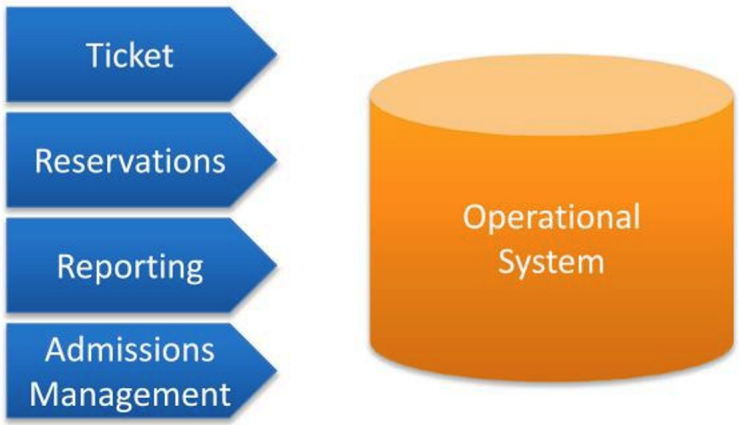
Definition

A Data Warehouse is a collection of support data for decision processes, which is:

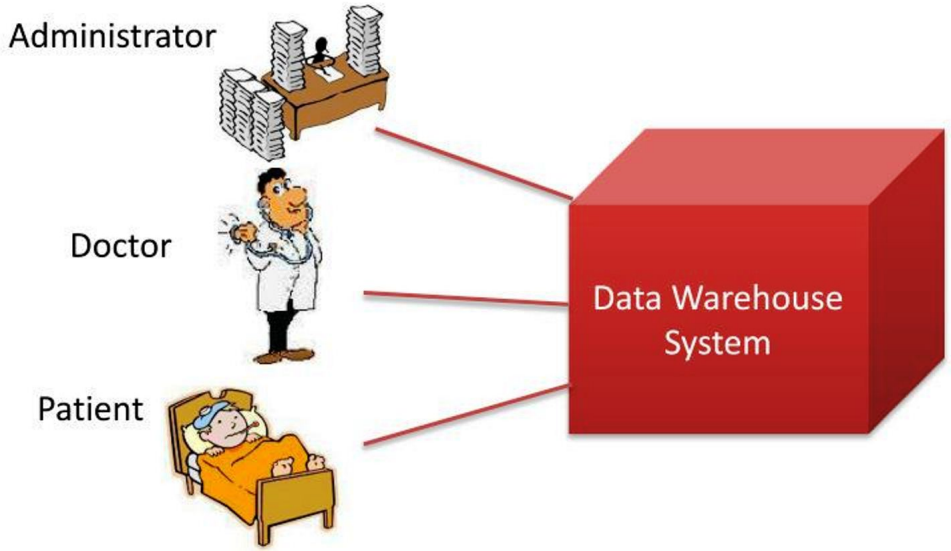
- subject-oriented;
- integrated and sound;
- representative of the temporal evolution;
- non-volatile.



Subject-Oriented



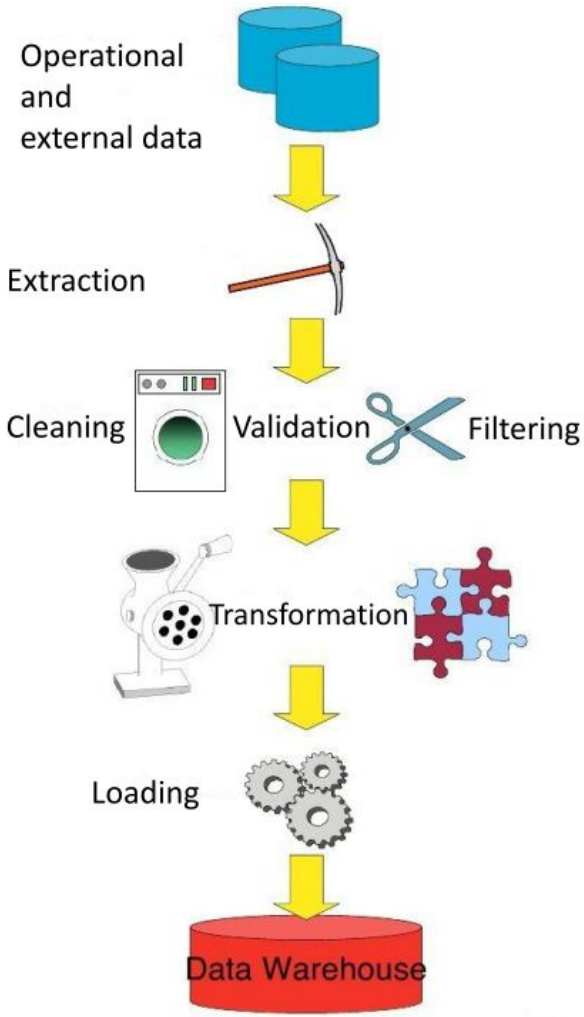
Application-Oriented



Subject-Oriented

Integrated and Sound

The DW relies on multiple sources of heterogeneous data ⇒ the goal is to return a unified vision.



Representative of the temporal evolution

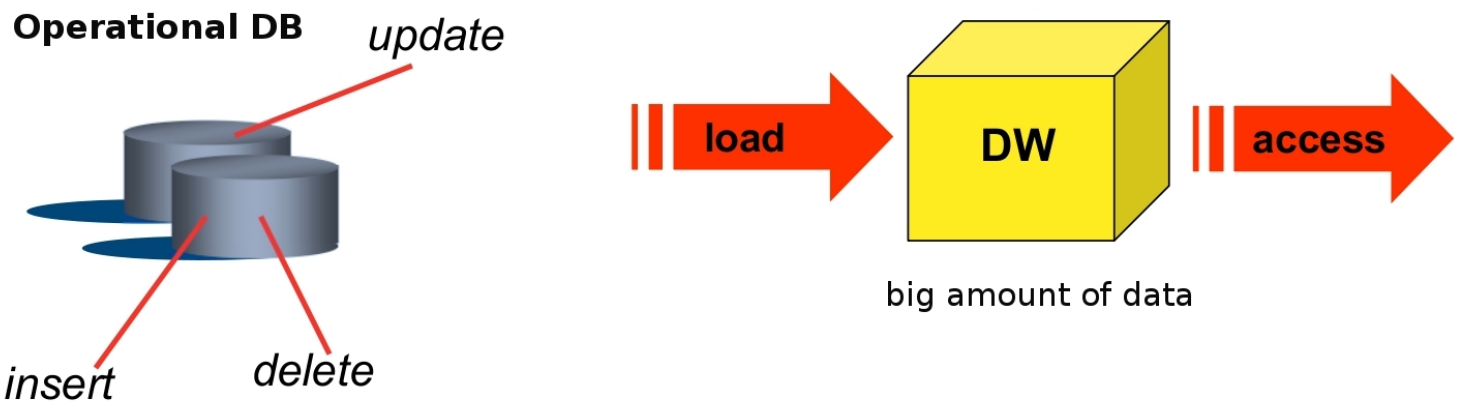
Operational DB

- limited historical content;
- time is not part of keys;
- data updates.

Data Warehouse

- rich historical content;
- time is part of keys;
- data cannot be updated/modified.

Non-volatile



In principle data are never deleted from the DW and updates are performed off-line $\Rightarrow$ read-only.

Summarizing

	Operational DB	Data Warehouse
Users	thousand	hundreds
Workload	default transactions	ad-hoc queries
Access	hundreds of records in reading and writing	millions of records especially in reading
Purpose	depends on the application	decision support
Data	basic, numeric and alphanumeric	synthesis, numeric
Integration	by application	by subject
Quality	integrity	consistency
Temporal coverage	only current data	current and historical data
Updates	continue	periodic
Model	normalized	denormalized and multidimensional
Development	cascade	iterative

Outline

ETL Tools

The role of Extraction,Transformation and Loading tools is to feed a single data source, detailed, comprehensive, and of high quality, which may in turn feed the DW (*Reconciliation*).

During the feeding process of the DW, reconciliation takes place:

- when the DW is populated for the first time;
- when the DW is periodically updated.

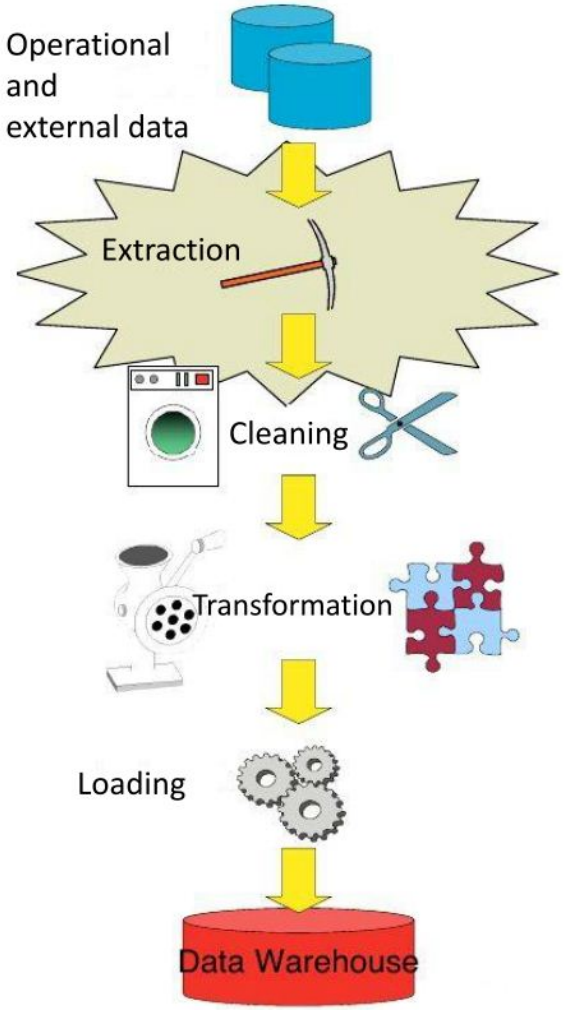
Stages of the reconciliation process:

1. extraction
2. cleaning
3. transformation
4. loading

Extraction

The relevant data are extracted from the sources. The choice on what data to extract is based on their quality.

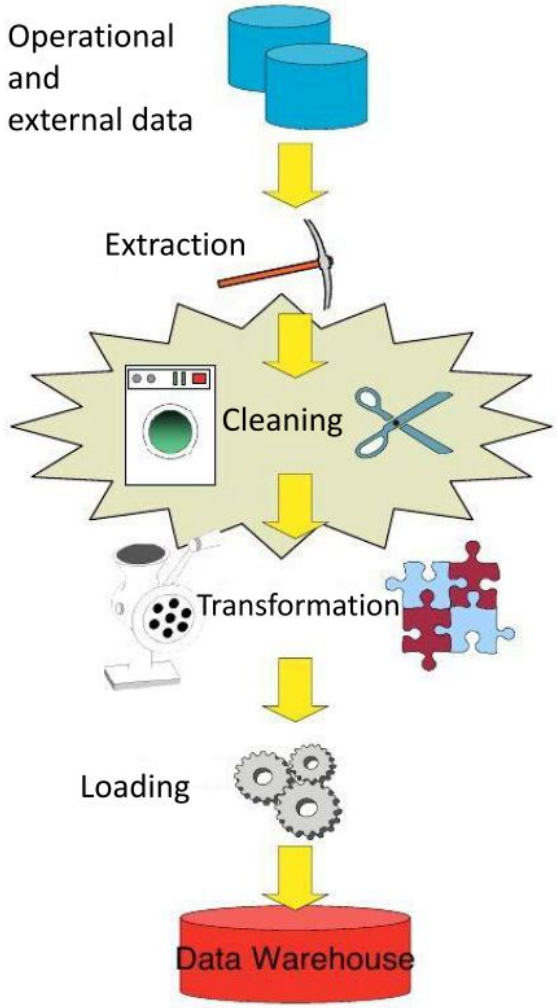
- Static extraction: when the DW is populated for the first time (snapshot of operational data).
- Incremental extraction: when the DW is periodically updated (captures the changes in the sources since the last update).



Cleaning

Improving the quality of the extracted data:

- duplicate data
- inconsistency between values
- missing data
- misuse of a field
- impossible or incorrect values
- inconsistent values due to different conventions adopted
- inconsistent values due to typing errors



Transformation

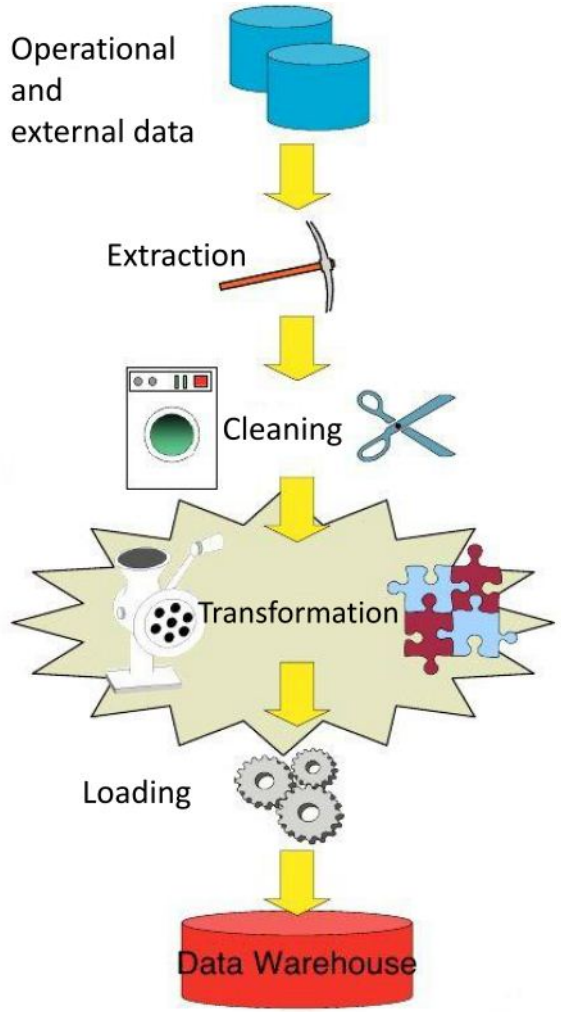
Converts the data into a uniform format.

Feeding of reconciled data:

- conversion and normalization: modify the format and the unit of measure to standardize data;
- matching: establishes correspondences between equivalent fields from different sources;
- selection: reduces the number of fields and records compared to the sources.

Feeding of DW:

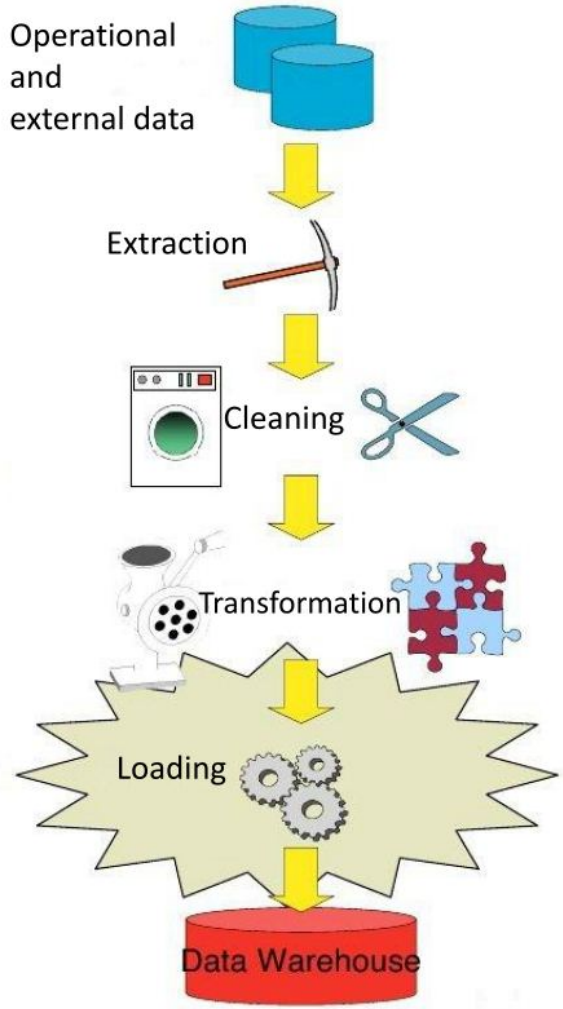
- denormalization: replaces the normalization;
- aggregation: makes appropriate summary of data.



Loading

Data loading on DW:

- Refresh: data are completely rewritten, previous data are replaced;
- Update: data are added to DW only when a change occurred in sources.



Outline

Multidimensional Model

The model allows one to represent and query data stored in the Data Warehouse.

- A Data Warehouse is usually built incrementally and is composed of one or more data marts.
- A Data Mart may be composed of several Cubes.

Facts of interest are represented in cubes, where:

- each cell of the cube contains numerical measures that quantify the fact;
- each axis of the cube represents a dimension of interest for the analysis;
- each dimension can be the root of a hierarchy of attributes used to aggregate data.

Admissions Cube

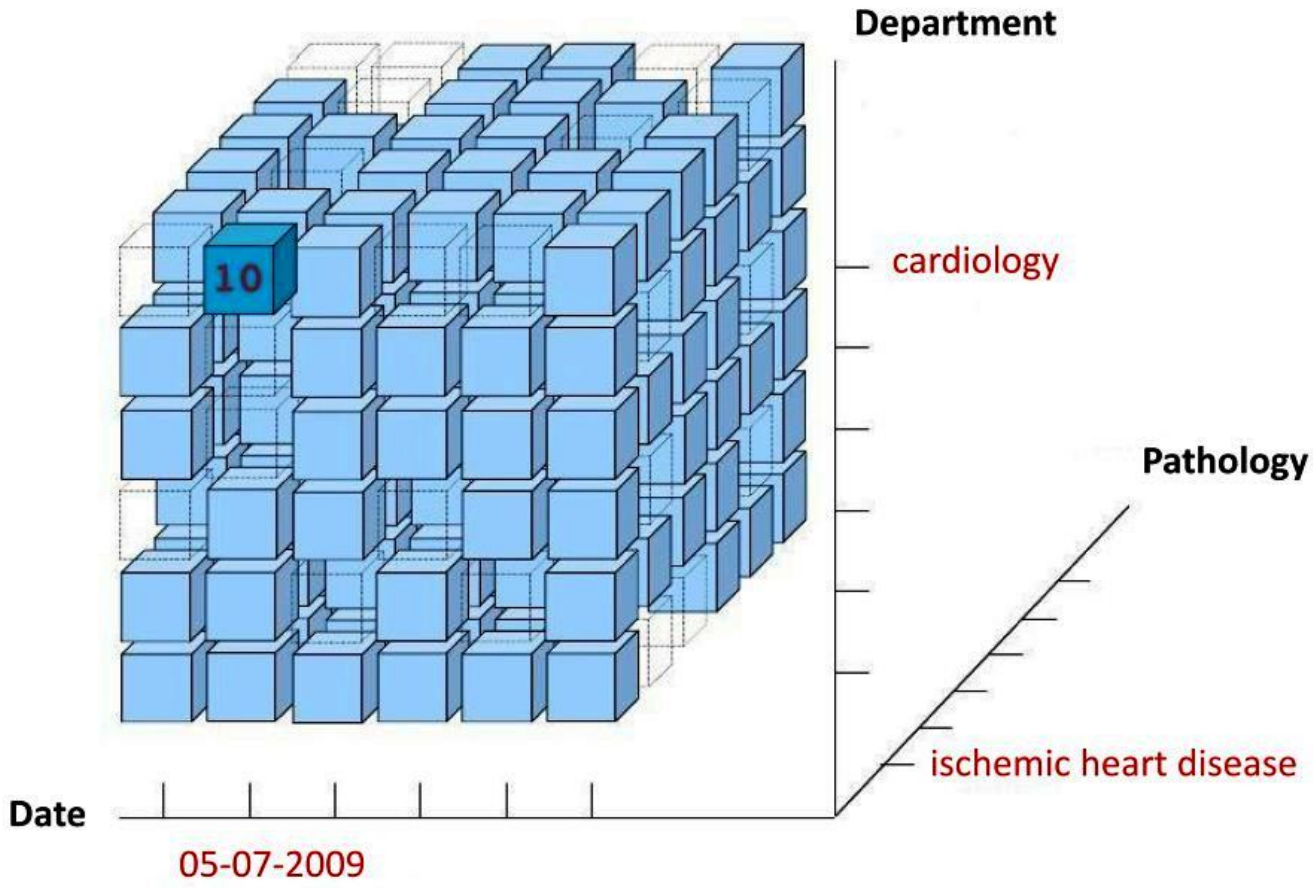
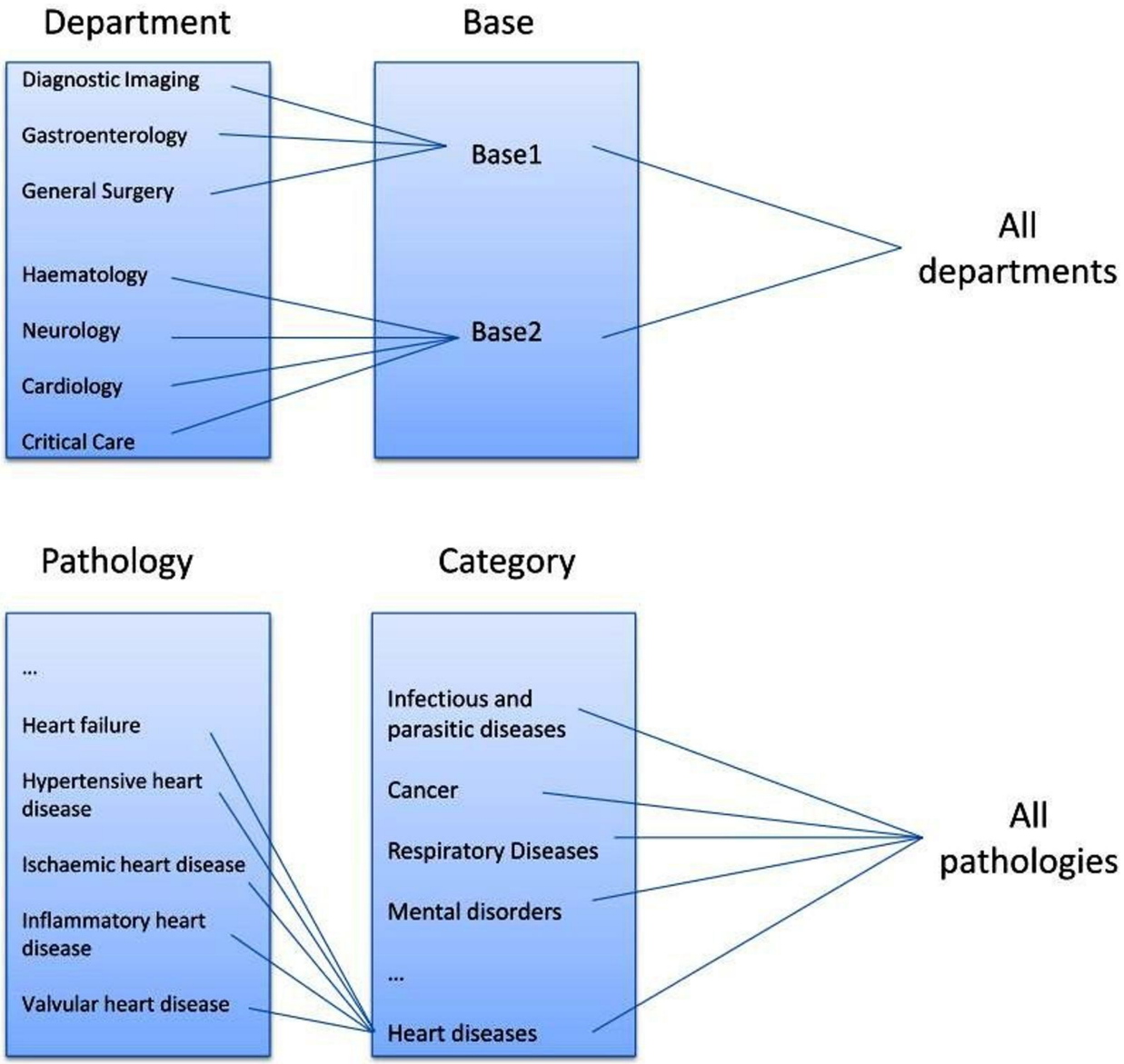
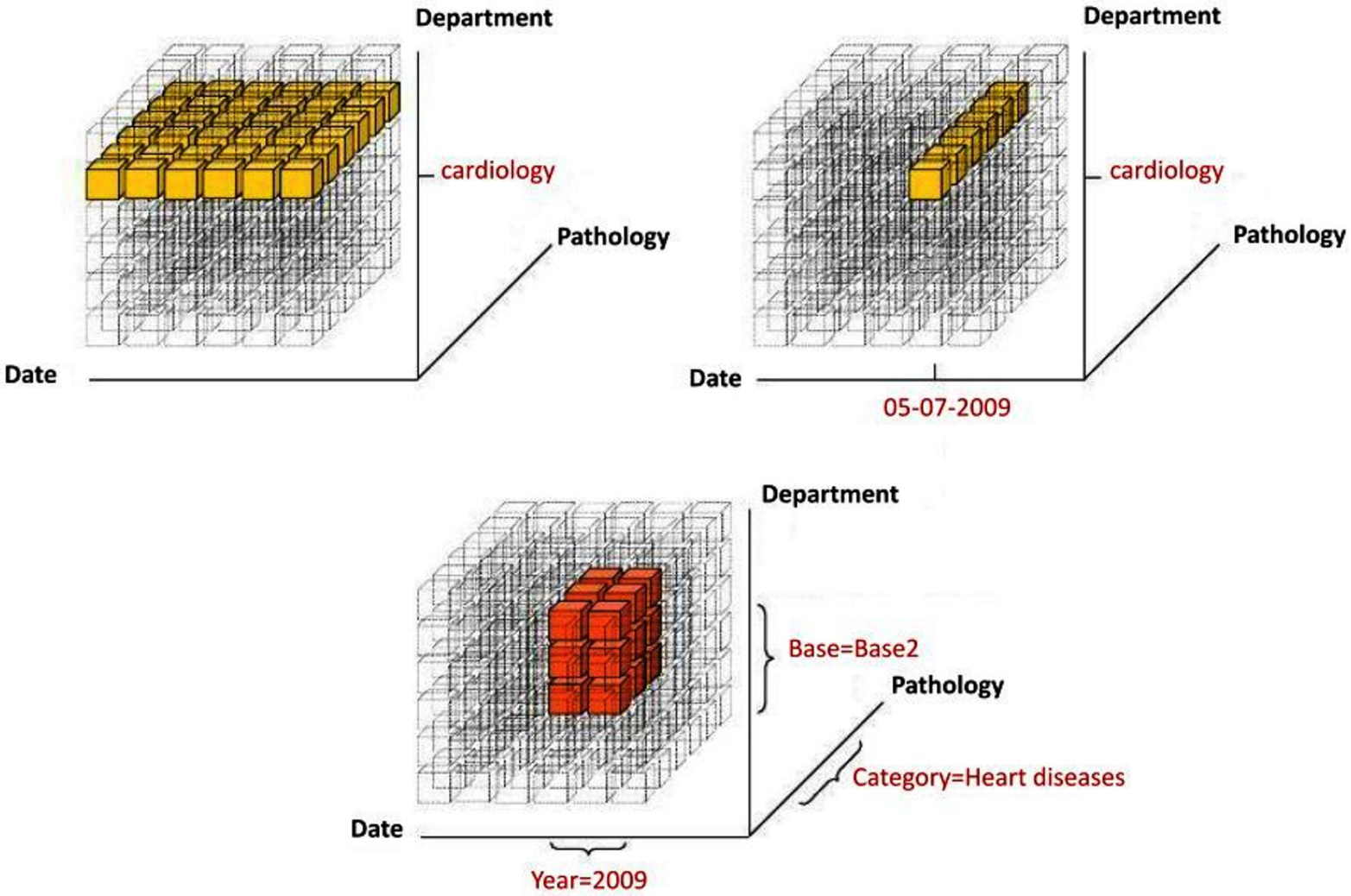


Figure: On 05/07/2009, 10 patients affected by ischemic heart disease were admitted to the cardiology department.

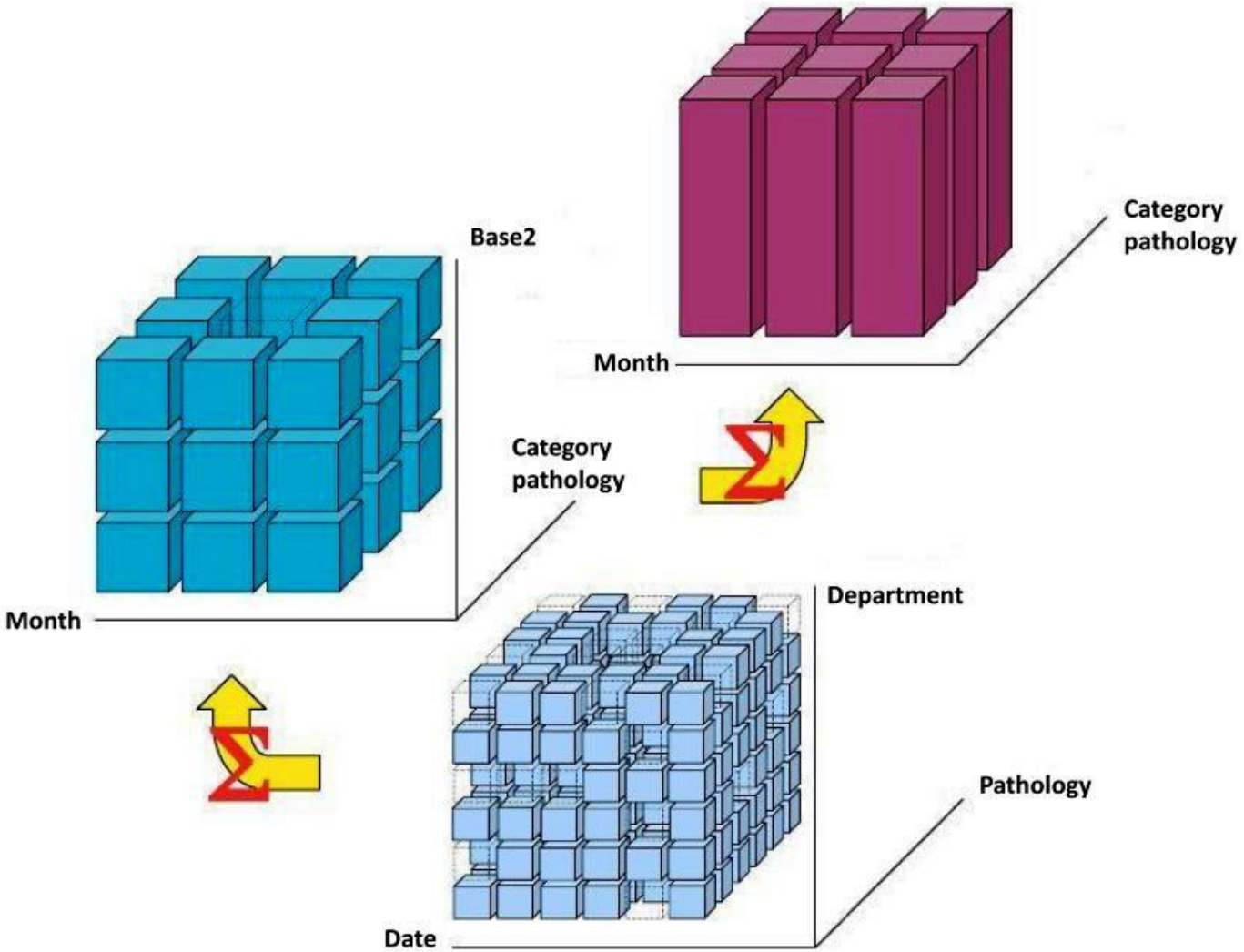
Hierarchies



Slicing and Dicing



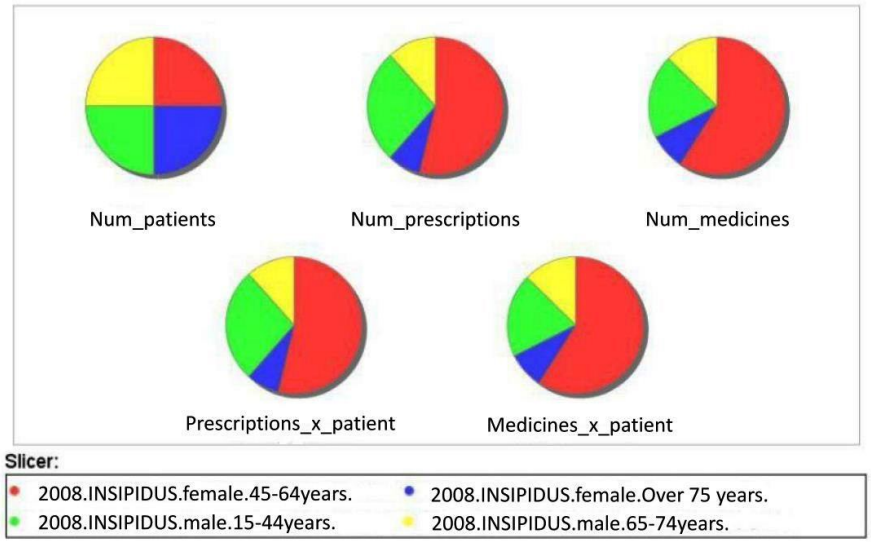
Roll-up and Drill-Down

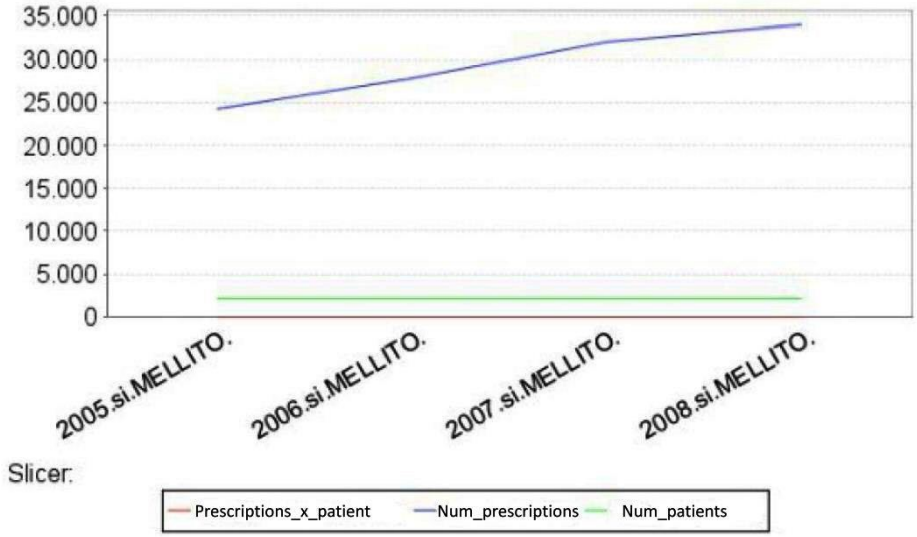


Outline

Data analysis techniques: Reporting

For users who periodically need to access to information with a fixed structure.





Data analysis techniques: OLAP

- OLAP users are able to actively build a complex analysis session in which each step is a consequence of previous results.
- Flexible interface.
- Easy to use and effective.

Definition

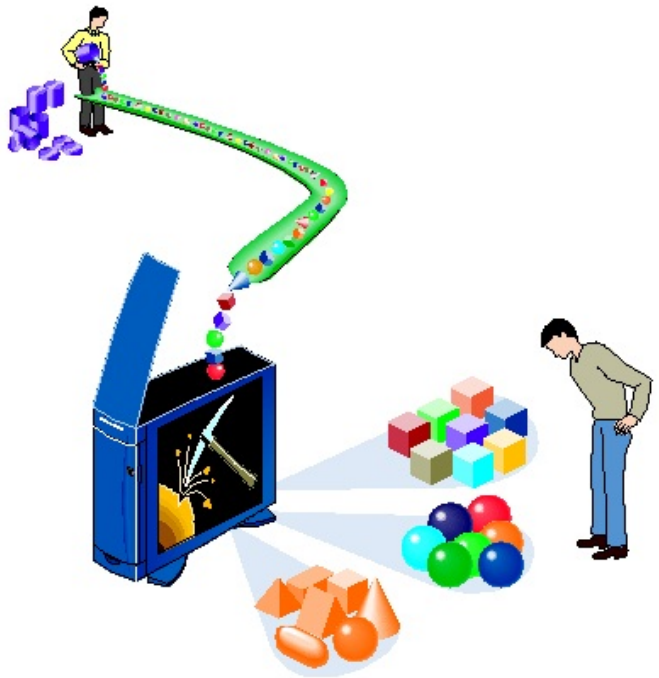
OLAP session: navigation path in the analysis cube.
It consists of an analysis of one or more facts of interest at different levels of detail. This path is composed by a sequence of queries often formulated by difference with respect to the previous query.

- Each step of the analysis session is made up by the application of an OLAP operator (Slice, Dice, Roll-up and Drill-down).
- The result is a multidimensional (sub)cube.

Data analysis techniques: Data Mining

Definition

Data mining is the process of extracting new knowledge from large data sets by combining methods from statistics and artificial intelligence with database management: clustering, time series analysis, what-if analysis, association rules, pattern recognition, probabilistic reasoning, classification.



Outline

Dimensional Fact Model - DFM

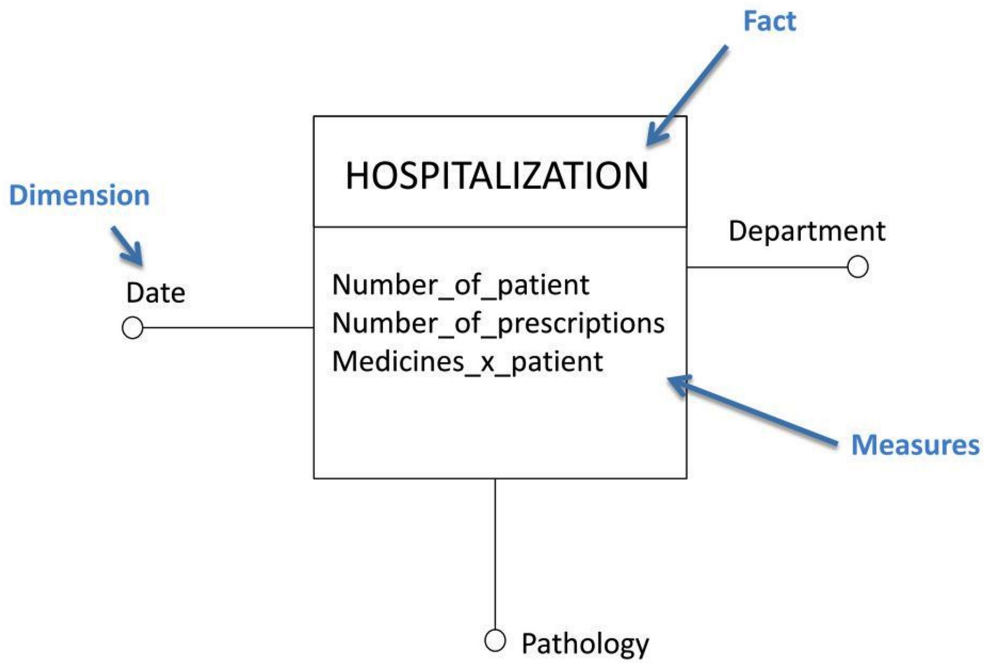
DFM is a graphical conceptual model for data marts, designed to:

- efficiently support the conceptual design;
- create an environment where the user can specify queries in a easy way;
- enable the dialogue between designer and end user to refine the specified requirements;
- create a stable platform to derive the design at the logical level;
- return expressive and unambiguous documentation.

The conceptual representation generated by DFM is a set of fact schemas. The basic elements modeled by fact schemas are: *facts*, *measures*, *dimensions* and *hierarchies*.

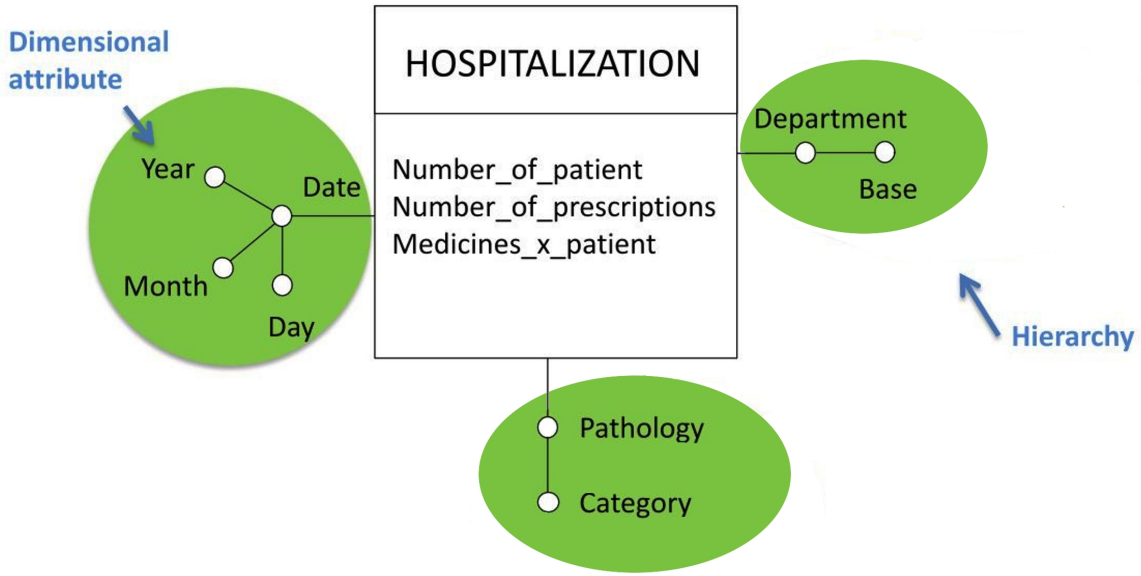
DFM: Basic Constructs

- Fact: concept of interest for decision making. It represent a set of events that occur in the considered (clinical/healthcare) domain and evolve over time.
- Measure: is a numeric property of a fact and describes a quantitative aspect of interest for the analysis.
- Dimension: is a property of a fact with a finite domain.



DFM: Basic Constructs

- Dimensional attributes: dimensions and any other related attribute.
- Hierarchy: directed tree whose nodes are dimensional attributes and whose edges represent associations many-to-one between pairs of attributes.



Events

Definition

A primary event is a particular occurrence of a fact, identified by a tuple consisting of a value for each dimension and each measure.

A secondary event is the aggregation of all corresponding primary events, according to the values of some dimensional attributes not being root of the hierarchy they belong to. Each secondary event is associated with a value for each measure.

Additivity

Aggregation requires to define a suitable operator to compose measure values of the primary events to calculate measure values to be associated with secondary events.

There are three categories of measures:

- flow measurements: it refer to a period after which they are assessed cumulatively;
- level measurements: are evaluated at particular time instants;
- unit measurements: are evaluated at particular time instants, but they are expressed in relative terms.

	Temporal hierarchies	Non-Temporal hierarchies
flow measurements	SUM, AVG, MIN, MAX	SUM, AVG, MIN, MAX
level measurements	AVG, MIN, MAX	SUM, AVG, MIN, MAX
unit measurements	AVG, MIN, MAX	AVG, MIN, MAX

Outline

Data Warehouses and Pharmacovigilance

- Clinical studies before a drug is put on the market are not able to guarantee the absence of side effects of commercial drugs
- The surveillance through spontaneous reporting for products already on the market can detect the majority of side effects of drugs
- The reporting of suspected adverse drug reactions feed the database of the World Health Vigibase.
- The European equivalent, EudraVigilance, was instituted in 2001.

Data Warehouses and Pharmacovigilance

- EudraVigilance and Vigibase, however, collect only a portion of the Italian reporting forms and their content.
- The Italian report forms were initially stored in the database of the GIF group and later in the National Network of Pharmacovigilance (RNF), with a partial overlap.

Data Warehouses and Pharmacovigilance

- There was the need to develop an integrated database for storing data previously described, along with a procedure for its feeding (continuous and periodic) and the tools to efficiently extract derived aggregated values.
- A data warehouse system has been developed for the analysis of adverse drug reactions reported in Italy.
- The project has been realized in collaboration with the Pharmacovigilance Regional Centre of the Veneto.

Data Warehouses and Pharmacovigilance

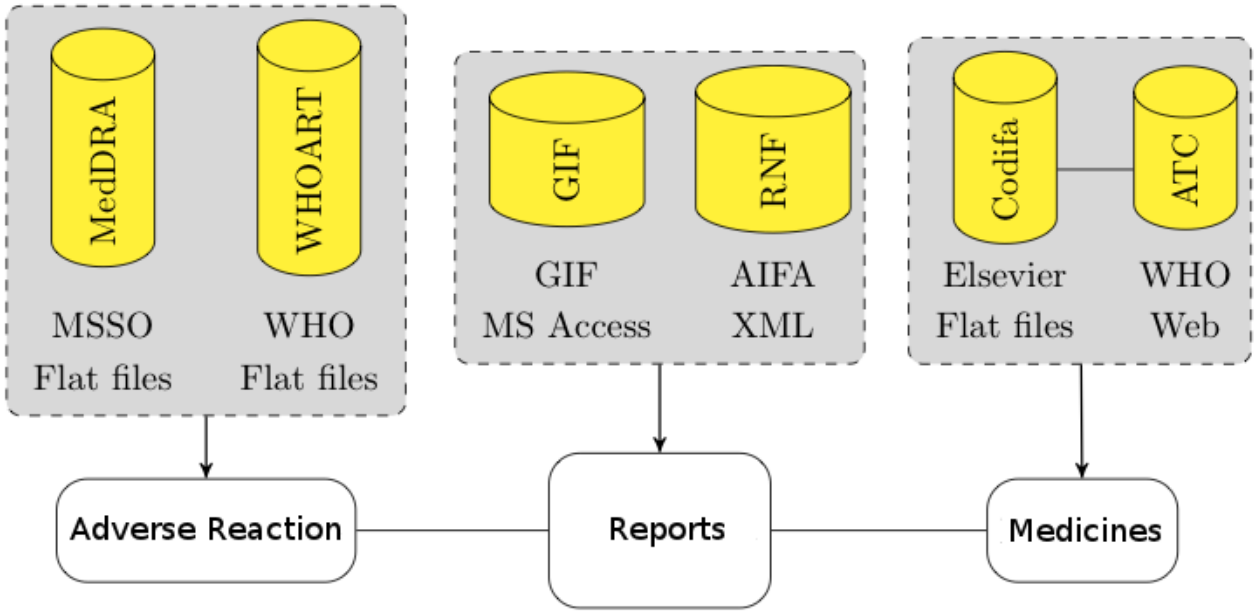
To develop such database, specific steps must performed (In this example we will focus only on the realization of the source level and the feeding level):

- Data Analysis
- Designing the system architecture
- Integration and feeding strategy
- Data Reconciliation
- ETL Phase

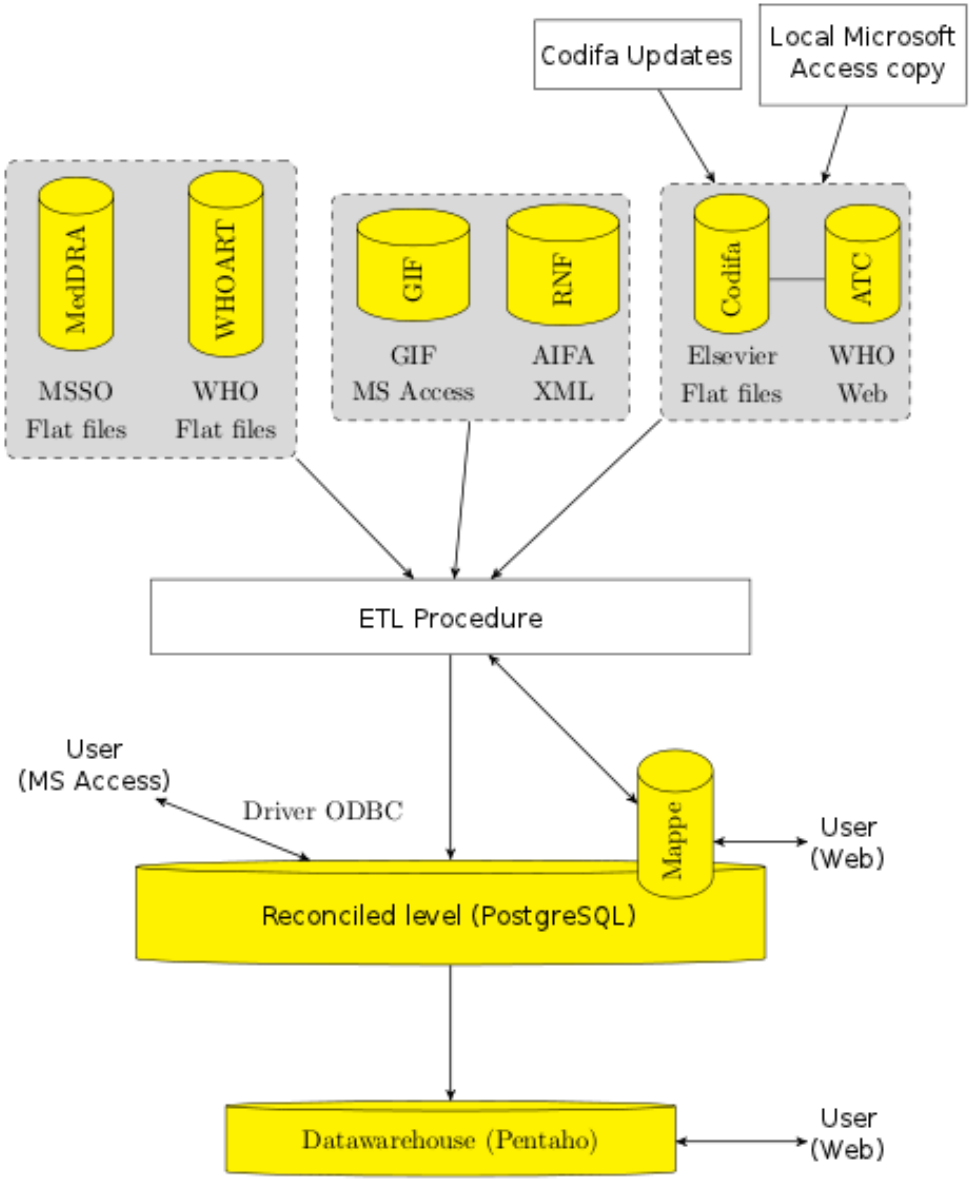
Data Analysis

Available databases

- drug database Codifa2000
- the databases for the MedDRA, WHOART, and ATC terminology
- GIF and RNF databases



The system architecture



Data Reconciliation

During the merge of the data from the GIF and RNF we found a problem in the choice of a global identifier for the report:

- The paper reporting forms are identified by a serial number and the initials of the region
- The RNF reporting forms are identified by a different sequence number managed by the Ministry of Health

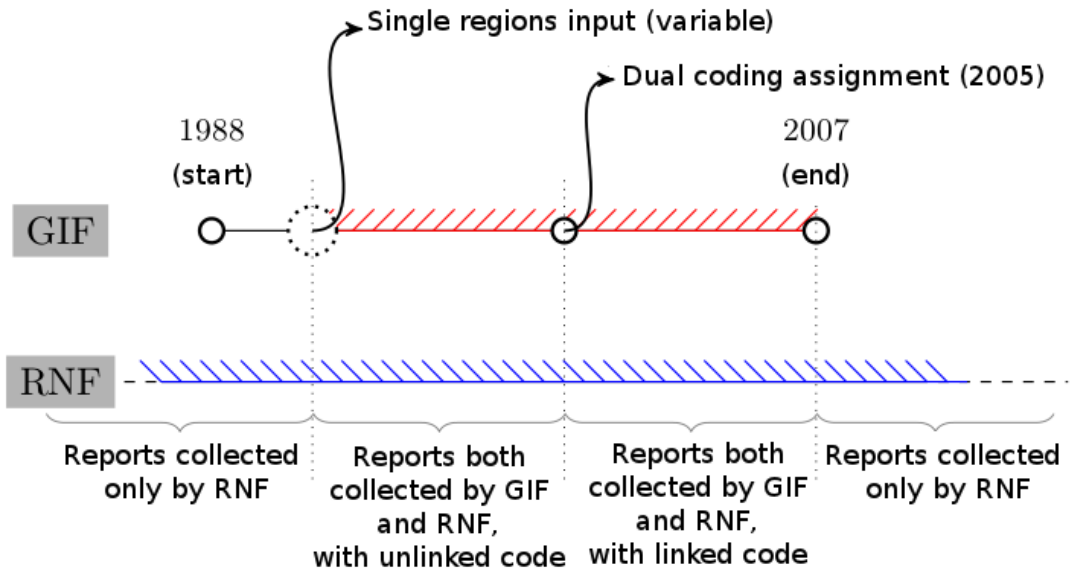
We decided to create a third identifier of 13 characters structured as follows:

GIF		
Local schema		Global schema
Region	Number	Code
VEN	26174	GIF-VEN-26174

RNF	
Local schema	Global schema
Number	Code
92143	AIFA-00092143

Data Reconciliation

Since the two archives have collected reporting forms filled in overlapping periods, it was also necessary to distinguish duplicates resulting from their merge. The reconciled level was then populated with all the reporting forms that are unique in the database of RNF as well as with those present in both databases, applying a selection to avoid duplicated forms.

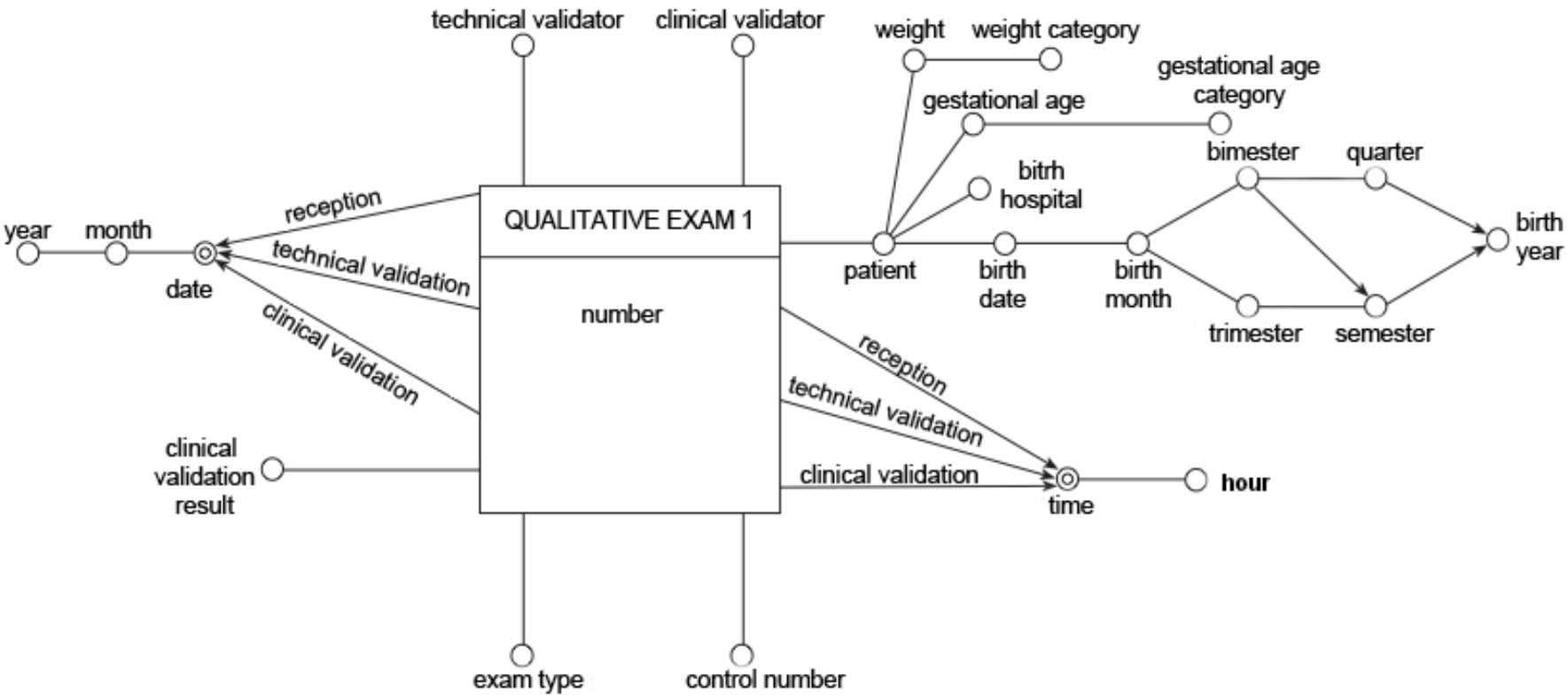


Data Warehouses and Neonatal Metabolic Diseases

- The Regional Centre for Neonatal Metabolic Diseases (CRMMN) of Verona performs newborn screening for major hereditary metabolic and endocrine diseases
  - galactosemia
  - PKU (Phenylketonuria syndrome)
  - Biotinidase deficiency
  - deficiency of glucose-6-phosphate-dehydrogenase (G6PD)
  - congenital hypothyroidism (IC)
  - Leucine or maple syrup urine disease (MSUD)
  - congenital adrenal hyperplasia (CAH)
- The screening is carried out on samples of blood taken from the heel of the newborn after 48 hours of birth

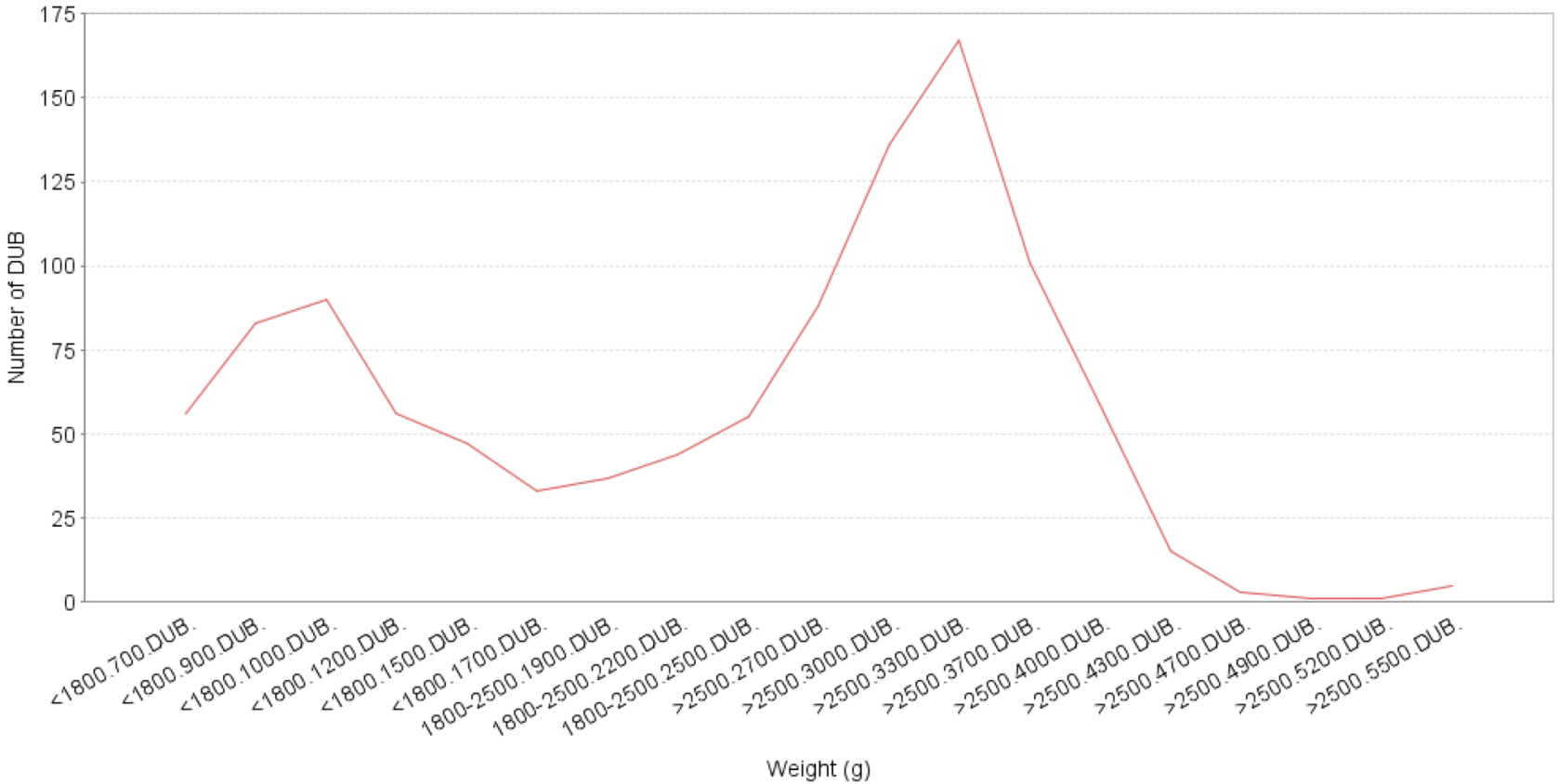
Qualitative assessment Fact schema

- The fact represented in this schema is the outcome of qualitative assessments, which records the result of the clinical validation of each quality examination carried out.
- The measures of interest are the number of times each outcome occurred of qualitative assessments (number of NO, VP, FN, DUB, etc.).



Doubt outcomes (DUB) grouped by weight

In order to identify differences in weight groups, possibly highlighting different risk levels, the data mart represents the number of tests that indicate suspect values, according to the weight of the newborn.



Total outcome number

We can see the total number of results by selecting only the dimensions date of birth, gestational age, type of examination, number of control and outcome. A total of 808,625 screening examinations has been performed.

					Measures
Birth Date	Gestational Age	Exam Type	Control Number	Outcome	• Number
⊞All BirthDates	⊞All Patiente.GestAges	⊞All ExamTypes	⊞All ControlNumbers	⊟All Outcomes	808.625
				AE	18
				CTOFF	126
				DEC	17
				DUB	3.208
				DUBtel	110
				FN	2
				II/III	1
				NO	803.323
				Non Noto	24
				PO	2
				RETEST	44
				VP	154
				X	40
				cdc	1.556

Total DUB recall for patient born in 2010

Selecting only the **DUB** results and focusing only on patients born in 2010, it can be noticed that there were 1,495 recalled.

					Measures
Birth Date	Gestational Age	Exam Type	Control Number	Outcome	• Number
⊞2010	⊞All Patiente.GestAges	⊟All ExamTypes	⊞All ControlNumbers	⊟All Outcomes	358.073
				DUB	1.495

Drill-down on the type of examination

By performing a drill down peration on the type of examination, it is possible to highlight how the quality tests carried out are grouped when the outcome is DUB.

					Measures
Birth Date	Gestational Age	Exam Type	Control Number	Outcome	• Number
⊞2010	⊞All Patiente.GestAges	⊟All ExamTypes	⊞All ControlNumbers	DUB	1.495
		BIOT	⊞All ControlNumbers	DUB	3
		G6PD	⊞All ControlNumbers	DUB	188
		GAL	⊞All ControlNumbers	DUB	6
		LEU	⊞All ControlNumbers	DUB	3
		N17OHPQUAL	⊞All ControlNumbers	DUB	234
		NT4QUAL	⊞All ControlNumbers	DUB	477
		NTSHSQual	⊞All ControlNumbers	DUB	549
		PKU	⊞All ControlNumbers	DUB	35

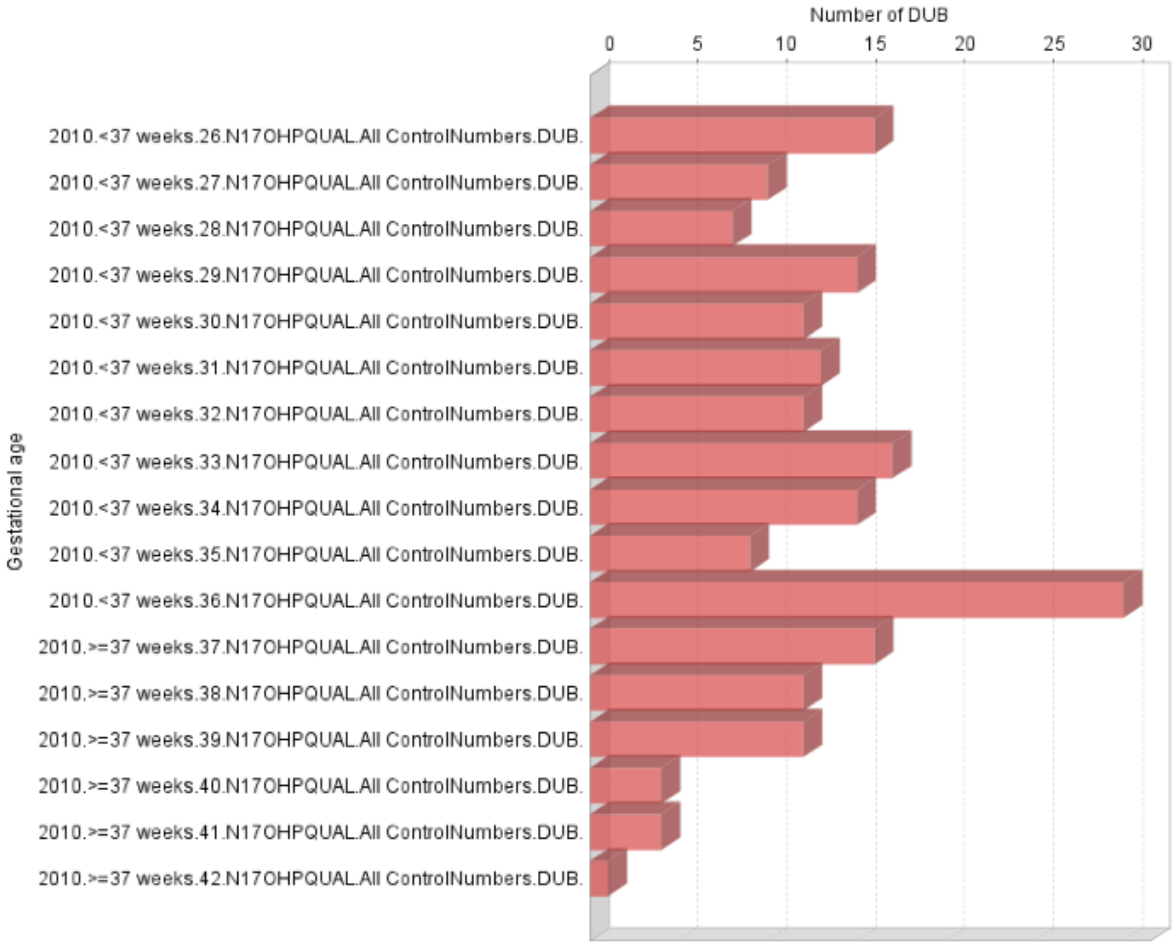
Focusing on the congenital adrenal hyperplasia

By selecting the exam type N170HPQUAL (qualitative exam for revealing the presence of CAH) and performing a drill-replace on the two members of the gestational age class level, we can see how the number of DUB for congenital adrenal hyperplasia (CAH) is related to the gestational age.

					Measures
Birth Date	Gestational Age	Exam Type	Control Number	Outcome	• Number
⊞2010	26	N17OHPQUAL	⊞All ControlNumbers	DUB	16
	27	N17OHPQUAL	⊞All ControlNumbers	DUB	10
	28	N17OHPQUAL	⊞All ControlNumbers	DUB	8
	29	N17OHPQUAL	⊞All ControlNumbers	DUB	15
	30	N17OHPQUAL	⊞All ControlNumbers	DUB	12
	31	N17OHPQUAL	⊞All ControlNumbers	DUB	13
	32	N17OHPQUAL	⊞All ControlNumbers	DUB	12
	33	N17OHPQUAL	⊞All ControlNumbers	DUB	17
	34	N17OHPQUAL	⊞All ControlNumbers	DUB	15
	35	N17OHPQUAL	⊞All ControlNumbers	DUB	9
	36	N17OHPQUAL	⊞All ControlNumbers	DUB	30
	37	N17OHPQUAL	⊞All ControlNumbers	DUB	16
	38	N17OHPQUAL	⊞All ControlNumbers	DUB	12
	39	N17OHPQUAL	⊞All ControlNumbers	DUB	12
	40	N17OHPQUAL	⊞All ControlNumbers	DUB	4
	41	N17OHPQUAL	⊞All ControlNumbers	DUB	4
	42	N17OHPQUAL	⊞All ControlNumbers	DUB	1

Focusing on the congenital adrenal hyperplasia

The peak of DUB for CAH coincides with 36 weeks of gestation, but fell considerably with higher gestational ages.

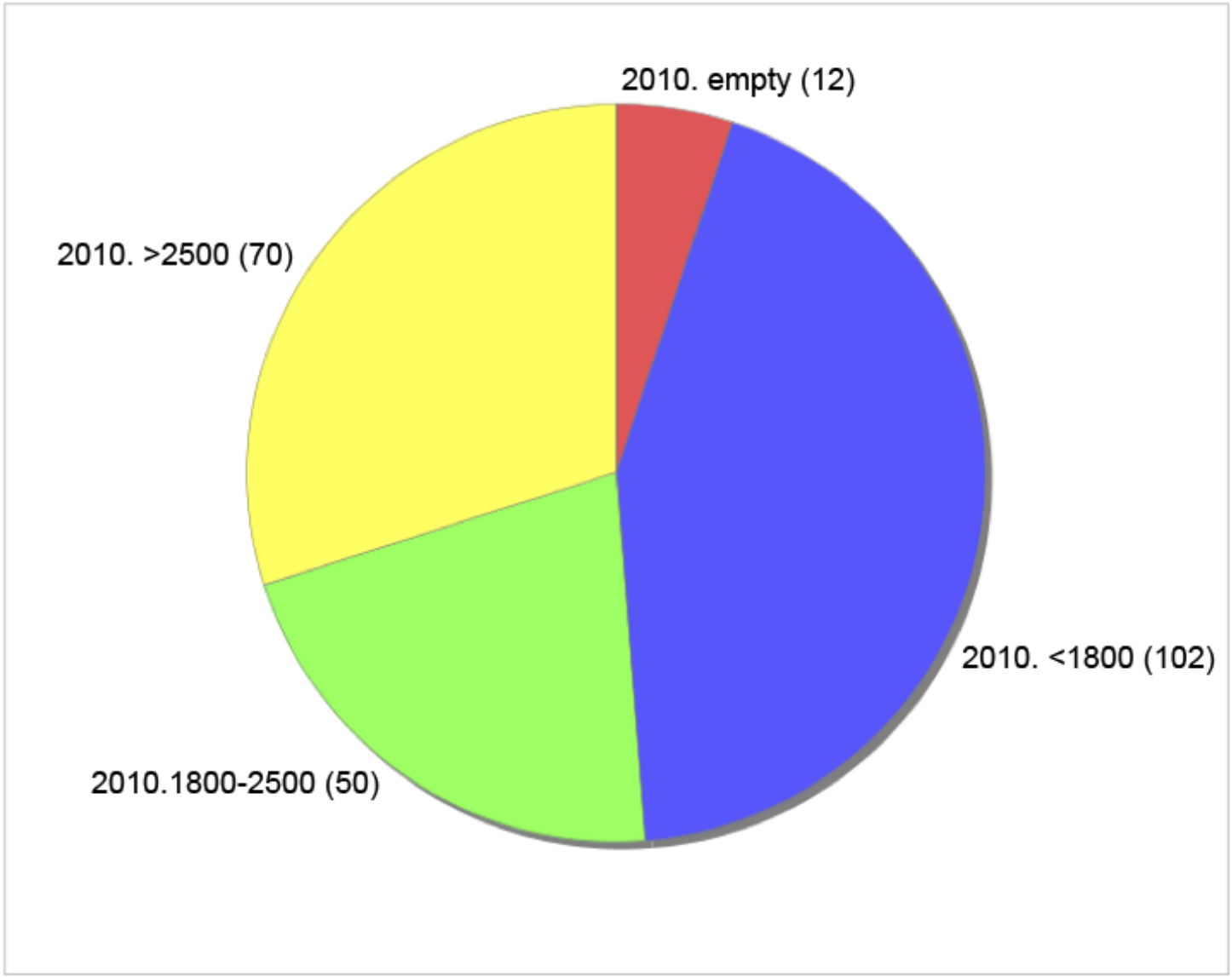


Using weight instead of age

By choosing to display the hierarchy of the patient's weight, instead of the gestational age, we can see the change in the number of DUB for CAH according to the weight category.

					Measures
Birth Date	Weight	Exam Type	Control Number	Outcome	• Number
⊞2010	⊞empty	N17OHPQUAL	⊞All ControlNumbers	DUB	12
	⊞<1800	N17OHPQUAL	⊞All ControlNumbers	DUB	102
	⊞1800-2500	N17OHPQUAL	⊞All ControlNumbers	DUB	50
	⊞>2500	N17OHPQUAL	⊞All ControlNumbers	DUB	70

Using weight instead of age

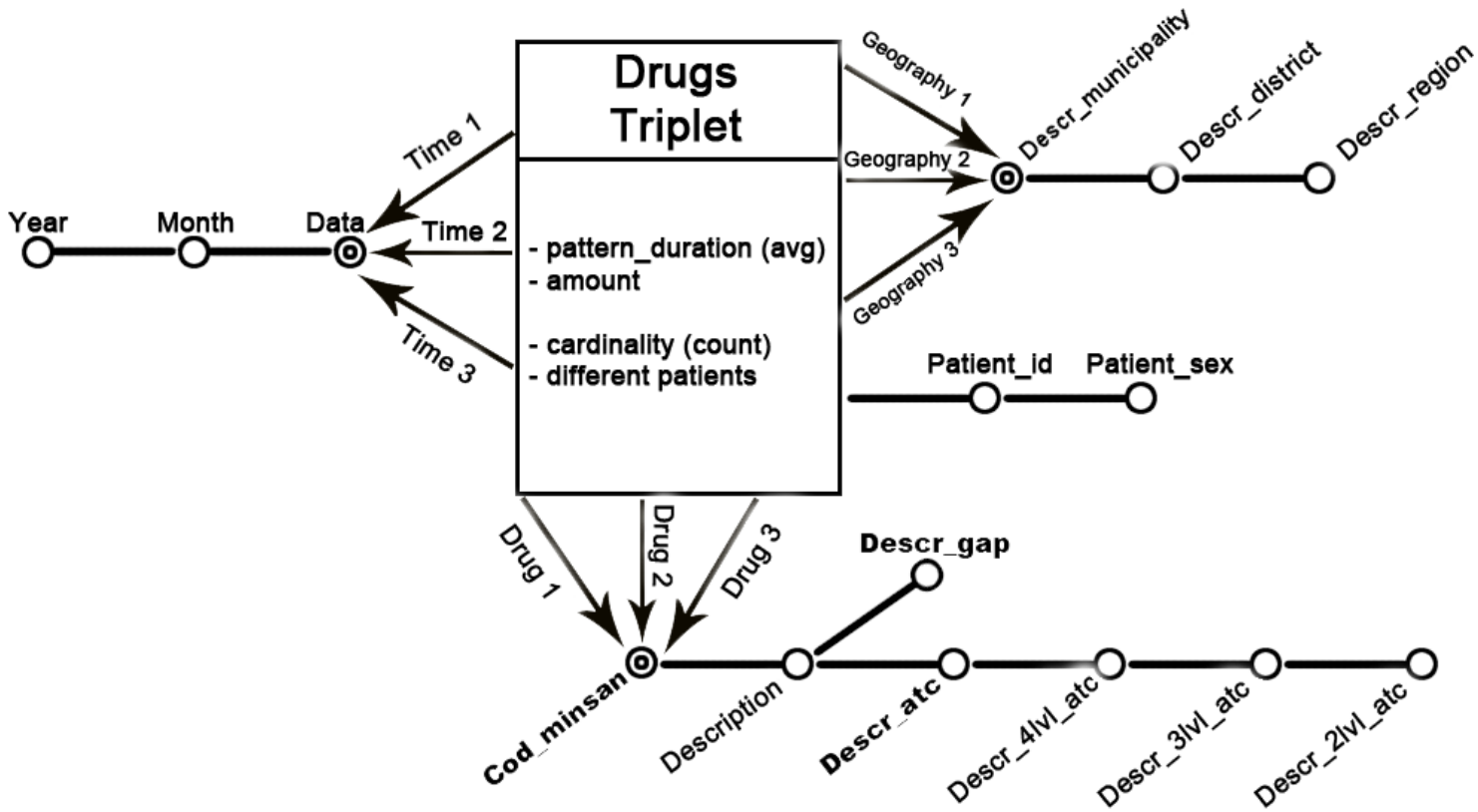


Data Warehouses and Drug Prescriptions

- We want to observe the pattern of three consecutive purchases of drugs, even the same product, for a patient;
- we consider a time span of 30 days as the maximum delay between the purchase of drug and the next one, and of 1 day as the minimum delay;
- The analysis focuses on the three most requested drugs for all patients in 2005.

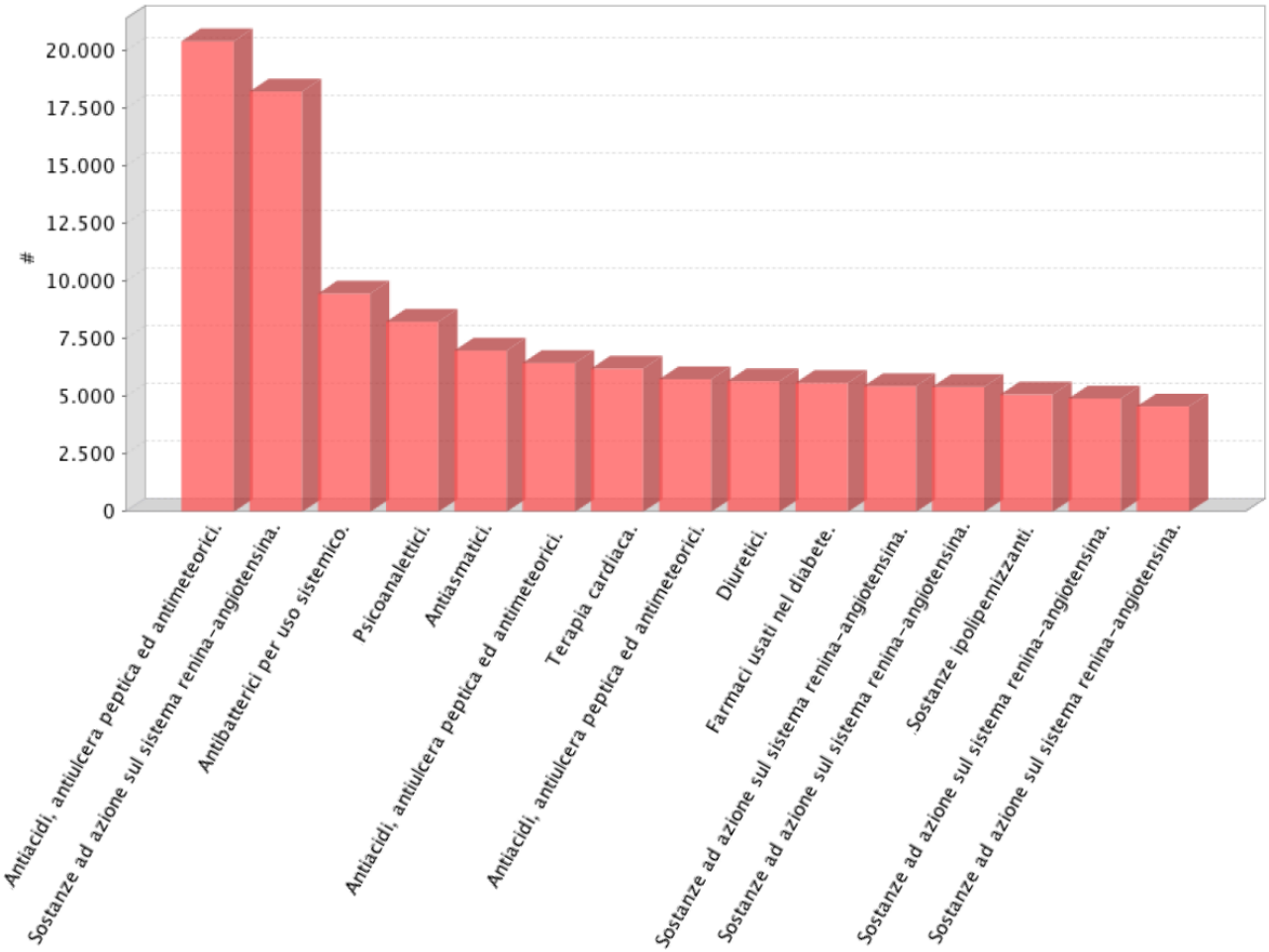
Drug patterns cube

This cube represents patterns of subsequent prescriptions for a given patient, within thirty days between a purchase of a drug and the next one.



Most requested drug triplets

Viewing the 15 consecutive prescriptions at the second level ATC that verifies the most requested pattern, we note that the one that contains three times "antacids, peptic ulcer and antimeteorics", exceeds any other triple with 20,384 requests.



Most requested drug triplets

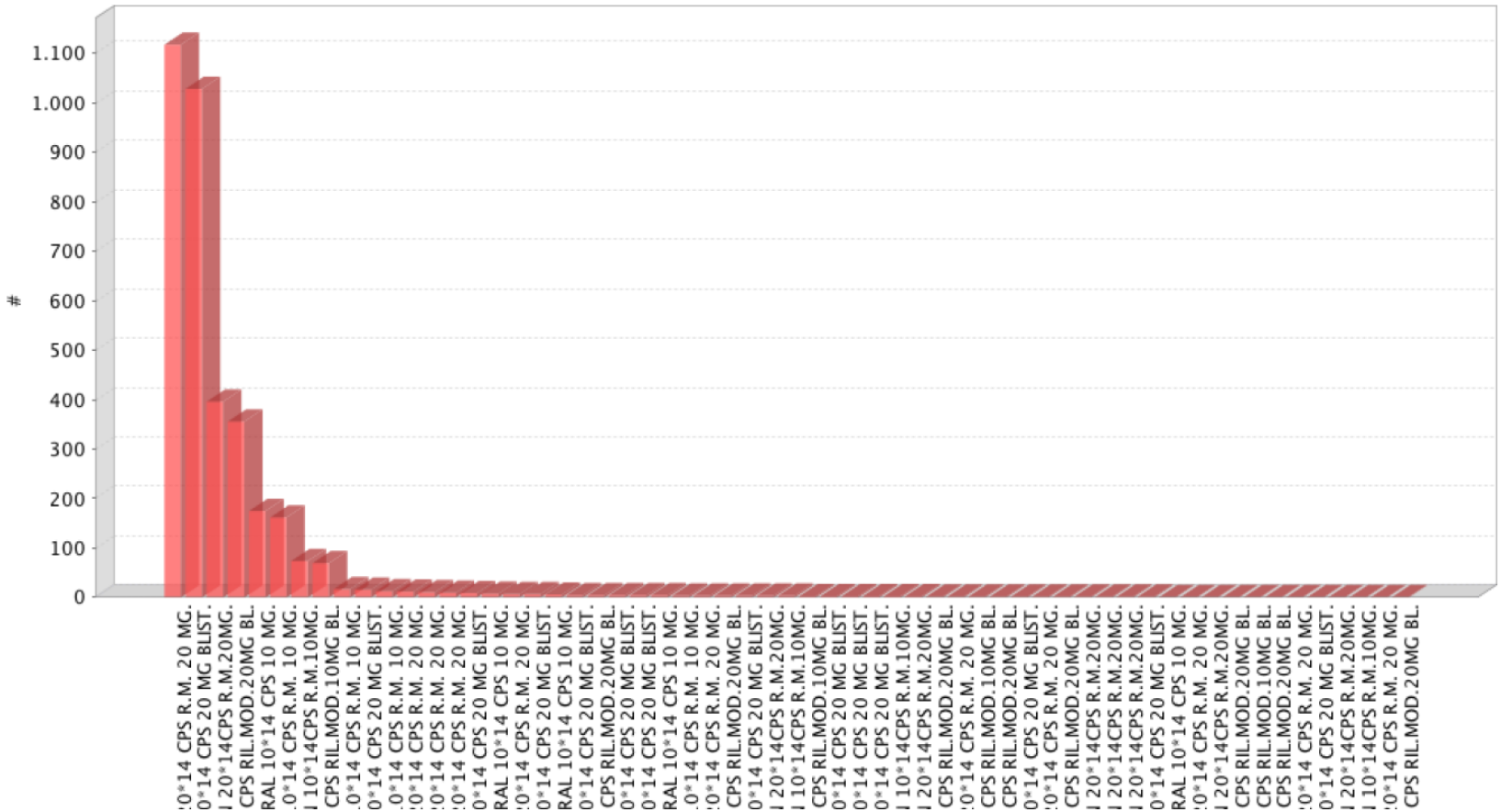
Selecting "antacids, peptic ulcer and antimeteorics" and moving along the hierarchy to the fourth ATC level, we note that the most popular triplet consists of three identical descriptions of the "acid pump inhibitors", with more than 12.000 instances.

Most requested drug triplets

T1.1 (omeprazol) is the triplet most purchased, at the description ATC level. We select only this triplet to reach the level description.

Most requested drug triplets

The drug "ANTRA 20*14 CPS R.M. 20 MG" at the Description level is in the most requested triplet, which has been requested 1,117 times.



Most requested drug triplets

The triplet was requested from 476 different patients, with an average duration of the pattern of 39.493 days, and the pharmaceutical company earned 170,523.72 euros.

			Measures			
Minsan Cod Drugs 1 atc	Minsan Cod Drugs 2 atc	Minsan Cod Drugs 3 atc	Cardinality	pattern_duration_avg	Amount	Different patients
028245090	028245090	028245090	1.117	39,493	170.523,72	476

Outline

Summary

- Having a huge amount of data makes it difficult to extract useful information.
- Data Warehouse System are the Decision Support System more established in the industrial world.
- As we have seen, this technology can be successfully used in the medical domain and allows:
  - Integrating different data sources, for a more complete and extended analysis
  - Finding unknown recurring pattern in the patients data
  - Discovering time patterns and improving the resource management