Information storage and management
to support
business decisions of organizations.

FROM DATA BASES TO DECISION SUPPORT DATA BASES

FACT In organizations, often the most important **decisions** are not based on fact (**informed decisions**), but on intuition and experience of managers.

FACT Organizations (companies) accumulate **large quantity of data**, that are often a resource scarcely used.

FACT Companies to compete **today** must use **data-intensive Business Intelligence techniques** to make better and timely business decisions.

FACT **Decision support information systems professionals** with a **business perspective** are needed to create company success and are rewarded by the job market.

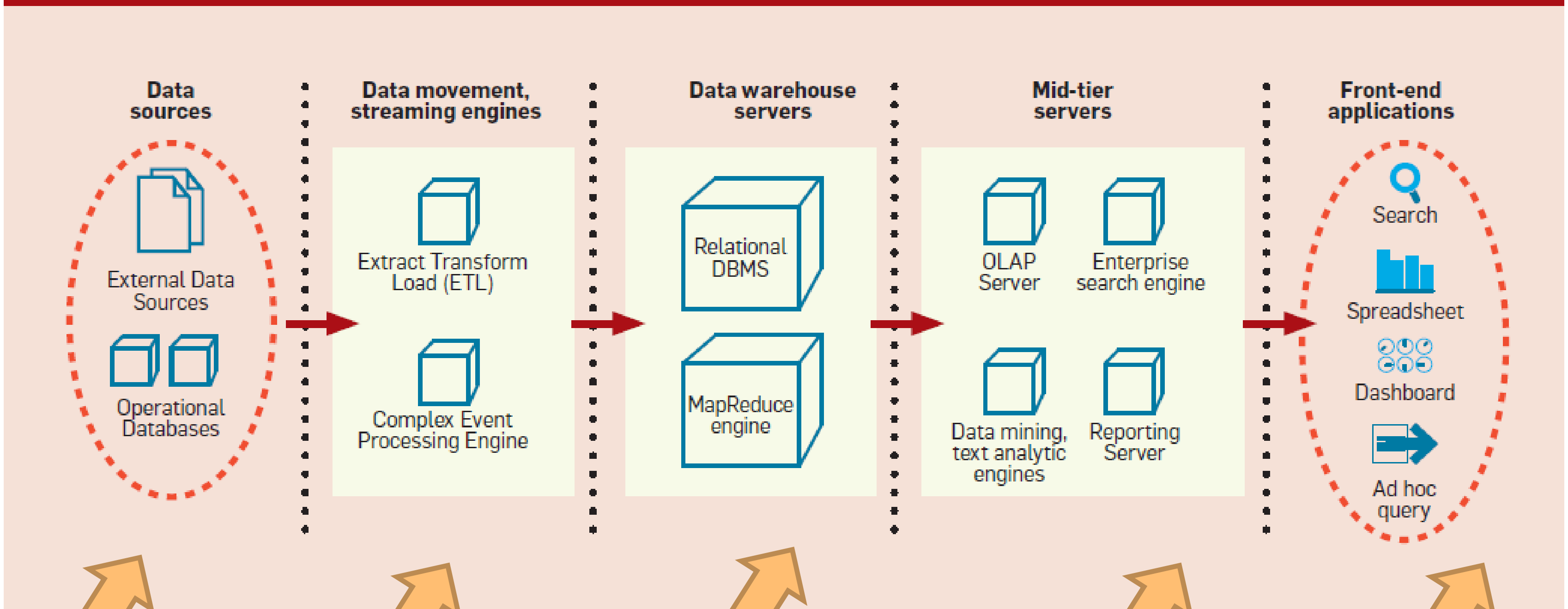
A set of methods and tools for interactive data analysis used to understand and analyze business performance in order to obtain useful information to **support unstructured decision making**.

The term **intelligence**...

... is used to mean search for something interesting, as in the **Intelligence Service**.

The design, implementation and use of a specific database, called **Data Warehouse (DW)**, to produce useful information to support decision-making with **Business Intelligence** applications

Figure 1. Typical business intelligence architecture.



Data access

ETL

DW Modeling

OLAP

Reporting

Operational System

- Data are organized in a **DB.**
- Data are managed by a **traditional DBMS**.
- The applications **are used to perform** structured business operational activities.

Decision Support System (DSS)

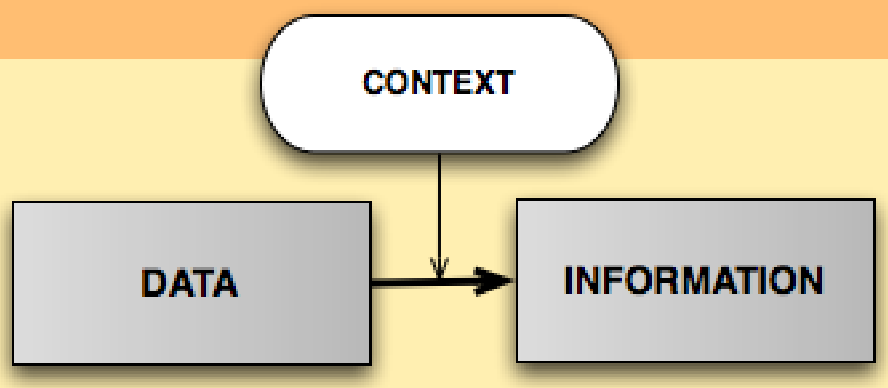
- Data are organized in a **separate specialized DB** (**Data Warehouse** (**DW**)).
- Data are managed by a **specialized DBMS**.
- The **Business Intelligence** applications, **are used to analyze data**.

Data

A representation of certain facts without context, which can be processed by computers.

442266	INF	2000	2003	Pisa
442277	TINF	2000	2004	Pisa
461176	IEA	2001	2003	Pisa
460076	TINF	2001	2003	Pisa
482299	INF	2002	2006	Pisa
481188	TINF	2002	2004	Pisa
441155	INF	2000	2002	Pisa
440033	TINF	2000	2002	Roma
498899	IEA	2003	2004	Bari
461178	INF	2001	2004	Bologna
…	…	…	…	…

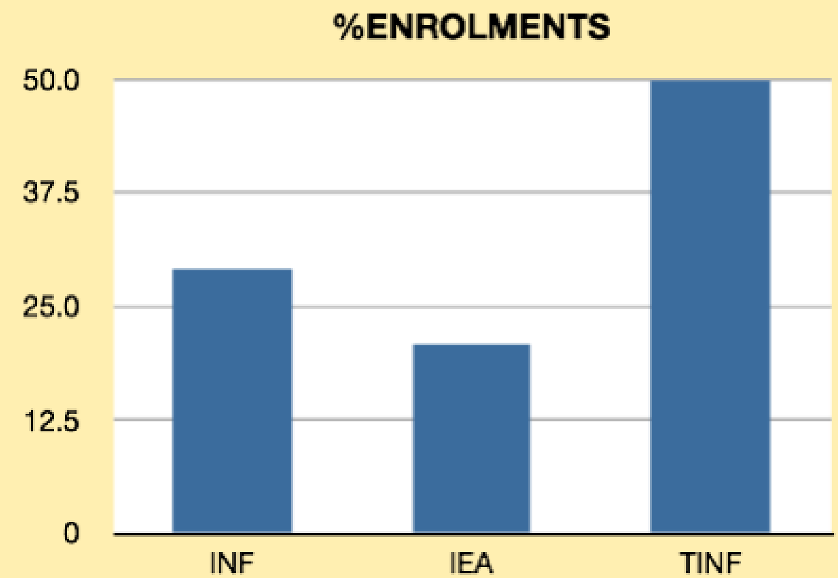
FROM DATA TO INFORMATION



Information

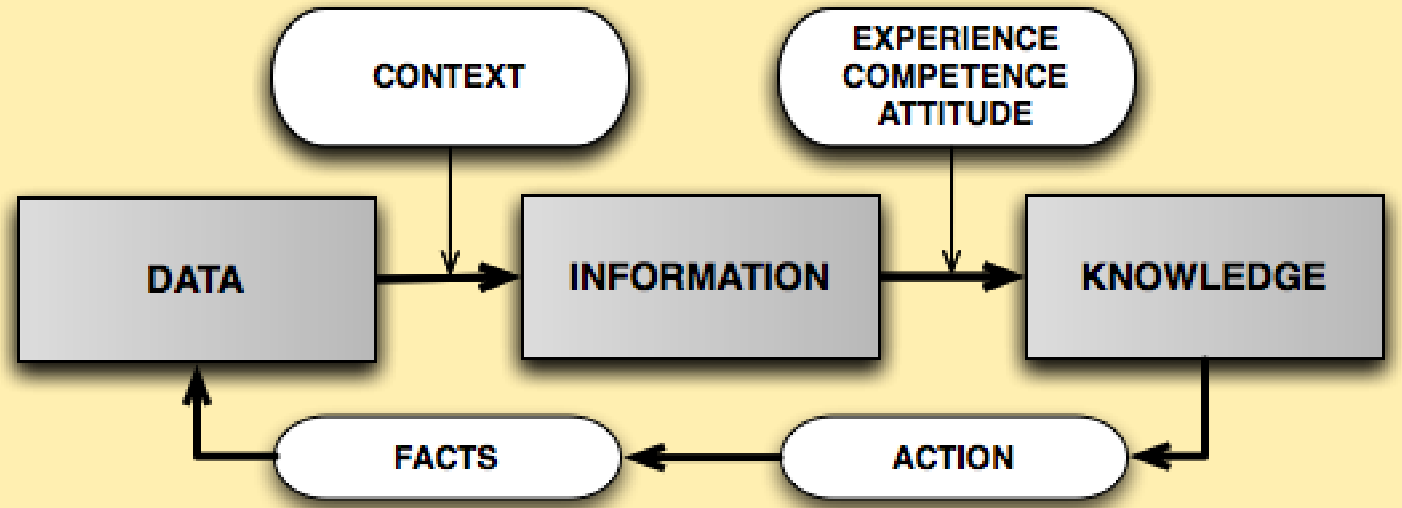
Data, or a condensed form of them, become information when they are interpreted in a certain context.

StudentN	Course	YearEnrol	YearDegree	FromUniv
442266	INF	2000	2003	Pisa
442277	TINF	2000	2004	Pisa
461176	IEA	2001	2003	Pisa
460076	TINF	2001	2003	Pisa
482299	INF	2002	2006	Pisa
481188	TINF	2002	2004	Pisa
441155	INF	2000	2002	Pisa
440033	TINF	2000	2002	Roma
498899	IEA	2003	2004	Bari
461178	INF	2001	2004	Bologna
…	…	…	…	…



Knowledge

Information become **knowledge** when **expand** the **recipient** capability of understanding the reality, and allow him to make new predictions, informed and effective decisions, and proper actions.



TYPES OF DATA SYNTHESIS

Reports: To find out what happened.

Multidimensional Data Analysis: To explore data interactively to look for useful information.

Exploratory Data Analysis: To discover useful models of data with **Data Mining** algorithms.

In what follows the attention will be on **Multidimensional Data Analysis**

Let us explore the sales data stored in the table

Sales(Product, Market, Date, Revenue)

For 2011, the total revenue, by semester.

Traditional Report

Revenue by Semester Year 2011	
Semester	**Revenue**
1	16 000
2	16 000
Total	**32 000**

Let us see if we can find more information with other business questions.

MULTIDIMENSIONAL DATA ANALYSIS (2)

UNIVERSITÀ DI PISA

For 2011, the total revenue, by market

Revenue by Market Year 2011	
Market	Revenue
M1	8 000
M2	8 000
M3	8 000
M4	8 000
Total	32 000

For 2011, the total revenue, by product

Revenue by Product Year 2011	
Product	Revenue
P1	8 000
P2	8 000
P3	8 000
P4	8 000
Total	32 000

MULTIDIMENSIONAL DATA ANALYSIS (3)

UNIVERSITÀ DI PISA

For 2011, the total revenue by semester, by product

Revenue by Semester, by Product Year 2011					
Semester	P1	P2	P3	P4	Total
1	4 000	4 000	4 000	4 000	16 000
2	4 000	4 000	4 000	4 000	16 000
Total	8 000	8 000	8 000	8 000	32 000

For 2011, the total revenue by semester, by market

Revenue by Semester, by Market Year 2011					
Semester	M1	M2	M3	M4	Total
1	4 000	4 000	4 000	4 000	16 000
2	4 000	4 000	4 000	4 000	16 000
Total	8 000	8 000	8 000	8 000	32 000

MULTIDIMENSIONAL DATA ANALYSIS (4)

For 2011, the total revenue
by semester, by product, by Market

OK, now we have got something interesting !

Revenue by Semester, by Product, by Market Year 2011						
Semester	Product	M1	M2	M3	M4	Total
1	P1			3 000	1 000	4 000
1	P2			1 000	3 000	4 000
1	P3	1 500	2 500			4 000
1	P4	2 500	1 500			4 000
	Total S1	4 000	4 000	4 000	4 000	16 000
2	P1	4 000				4 000
2	P2		4 000			4 000
2	P3			1 500	2 500	4 000
2	P4			2 500	1 500	4 000
	Total S2	4 000	4 000	4 000	4 000	16 000
Total		8 000	8 000	8 000	8 000	32 000

The result must be well visualized...

ANOTHER EXAMPLE

	Jan 2006	Feb 2006	Mar 2006	Apr 2006	May 2006	Jun 2006	Jul 2006	Aug 2006	Sep 2006	Oct 2006	Nov 2006	Dec 2006
Revenue	$688,588	$717,238	$750,969	$740,802	$728,844	$819,096	$818,760	$730,322	$836,197	$750,702	$818,315	$984,378
Revenue Forecast	$ 716,132	$ 631,169	$ 811,047	$ 788,296	$ 665,454	$ 728,995	$ 874,474	$ 691,552	$ 731,771	$ 623,083	$ 743,496	$ 1,000,201
Variance	(27,544)	86,069	(60,078)	(47,494)	63,389	90,101	(55,714)	38,770	104,427	127,619	74,819	(15,823)
Revenue Growth %	22.1%	37.4%	16.2%	35.1%	32.1%	44.2%	32.3%	7.8%	21.3%	9.9%	5.9%	20.0%

- **What** is a Data Warehouse (DW)
- **What** do we model in a DW
- **How** do we implement a DW
- **How** do we make multidimensional analysis

DEFINITION

A **DW** is a **decision support database** with **historical**, **nonvolatile data**, pulled together primarily from operational business systems, structured and tuned **to facilitate analysis of the performance of key business processes**, **worthy of improvement**.

The first definition of data warehouse was provided by William Inmon in 1990.

A **DW** is a **specialized database**

- **static (non volatile)**,
- **with integrated data** from different data sources,
- **organized to analyze subjects of interest**,
- **with historical data**,
- **used to produce summarized data to support decision-making processes**.

WHY SEPARATE DB AN DW?

To promote the high performance of both systems

- Special data organization, and implementation techniques are needed to support multidimensional OLAP analysis.
- Complex data analysis would degrade performance of operational DBMS.

The systems have different structures, contents, and uses of the data

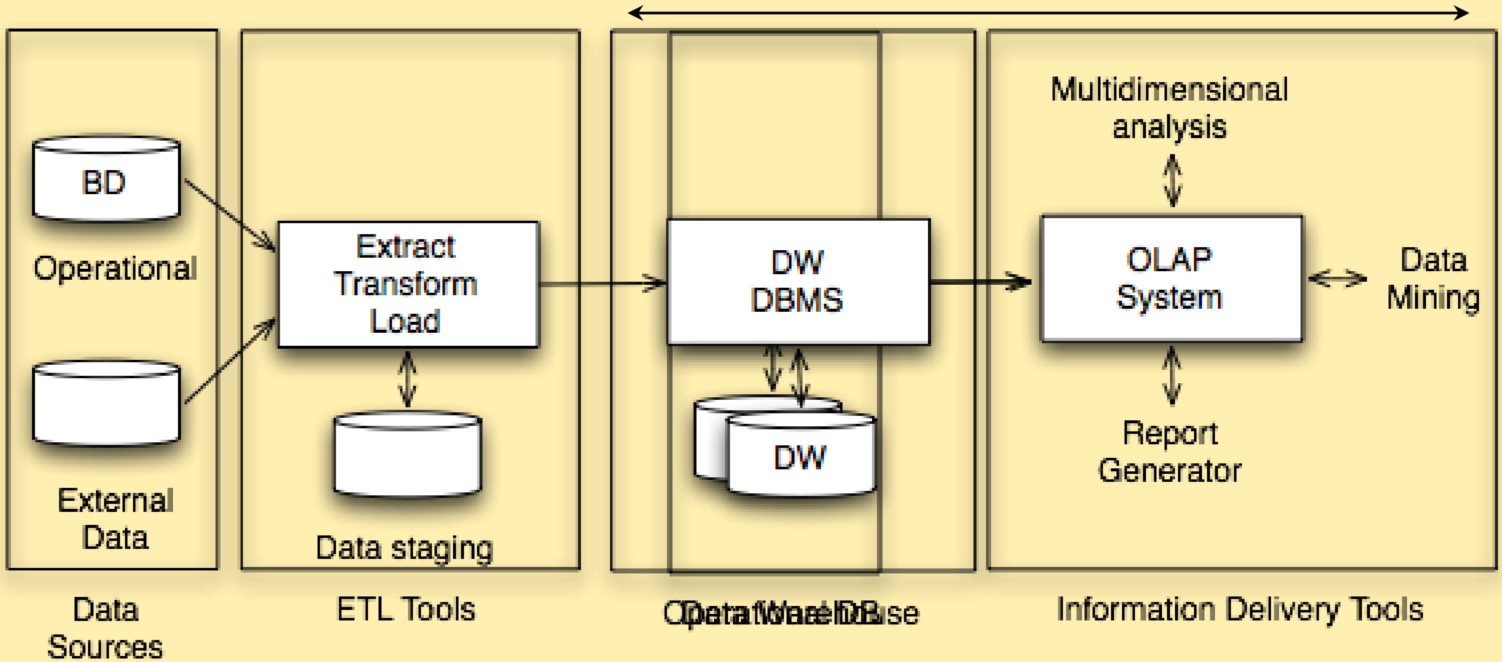
- Decision support requires historical data which operational DBs do not typically maintain.
- DS requires aggregation of data from heterogeneous sources: operational DBs, external sources.
- Different sources typically use inconsistent data representations, codes and formats which have to be reconciled.

Data warehousing is the process to bring data from operational (OLTP) sources into a single data warehouse for (OLAP) analysis with Business Intelligence applications.

DATA WAREHOUSING ARCHITECTURES

UNIVERSITÀ DI PISA

DSS



+ Data Visualization Tools

Single-Layer Architecture

Three-Layer Architecture

Business Intelligence

OLAP (On Line Analytical Processing)

WHAT IS MODELED IN A DW?

UNIVERSITÀ DI PISA

Managers think about a business process in terms of

facts, A **fact** is an observation of the performance of a business process (the subject of analysis) (e.g. the sales made into a period of time)

measures, The **measures** are numerical attributes of a fact (e.g. qty, revenue, etc),

... which are useless without a **context**.

dimensions, The **dimensions** give facts their **context.** In general a dimension is described by a **set of attributes**, otherwise is called **degenerate**. (e.g. sales revenue by **product** category, by month **time**, and by city **market**).

and **hierarchies**, The **attributes** of a dimension may be related via a **hierarchy** of relationships (e.g. a month is related to the quarter and the year attributes).

FACTS ANALYSIS

Managers are interested in aggregate data: the sum, average minimum, maximum, ..., of measures of data groups with equal values of some dimensions or dimensional attributes.

Metrics and **Key Performance Indicators** (KPI)

Total sales revenue, by products.

Total revenue, by Product (SQL ?)

SALES

Product	Store	Date	Revenue
p1	m1	d1	120
p2	m1	d1	110
p1	m3	d1	500
p2	m2	d1	800
p1	m1	d2	400
p1	m2	d2	300

```
SELECT    Product
          , SUM(Revenue) AS TotalRevenue
FROM      Sales
GROUP BY  Product ;
```

Product	Store	Date	Revenue
p1	m1	d1	120
p1	m3	d1	500
p1	m1	d2	400
p1	m2	d2	300
p2	m1	d1	110
p2	m2	d1	800

Product	Total Revenue
p1	1320
p2	910

Managers think in term of business dimensions to analyze the data and produces requirements

Total revenue by Product.
Total revenue by Product, by Market

Requirements for the design

Revenue by Product	
Product	Revenue (€)
P1	130
P2	910

Revenue by Product and Market		
Product	Market	Revenue (€)
P1	M1	520
	M2	300
	M3	500
P2	M1	110
	M2	800

FACTS ANALYSIS WITH SUBTOTALS

Revenue by Product and Market

Product	Market	Revenue (€)
P1	M1	520
	M2	300
	M3	500
P1	**Total**	**1 320**
P2	M1	110
	M2	800
P2	**Total**	**910**
Total		**2 230**

Group by Product, Market

Group by Product

No Group by

This result can be computed with a particular extension of SQL

FACTS ANALYSIS

Managers analyse measure aggregates by business dimensions, and then in various levels of details, by exploiting **dimensional attributes hierarchies.**

Total Revenue, by Month

Total Revenue, by Quarter

Total Revenue, by Year

Example with Sales of Year 2010

A dimensional attributes hierarchy models **attributes dependency**, i.e. a **functional dependency** between attributes, using the relational model terminology.

Definition 8.1 *Functional Dependency*

Given a relation schema R and X, Y subsets of attributes of R, a functional dependency $X \rightarrow Y$ (X *determines* Y) is a constraint that specifies that for every possible instance r of R and for any two tuples $t_1, t_2 \in r$, $t_1[X] = t_2[X]$ implies $t_1[Y] = t_2[Y]$.

For example, the dimension **Date** has attributes **Month**, **Quarter**, **Year**. Can we define a **dimensional hierarchy** among them?

Month → **Quarter** → **Year**

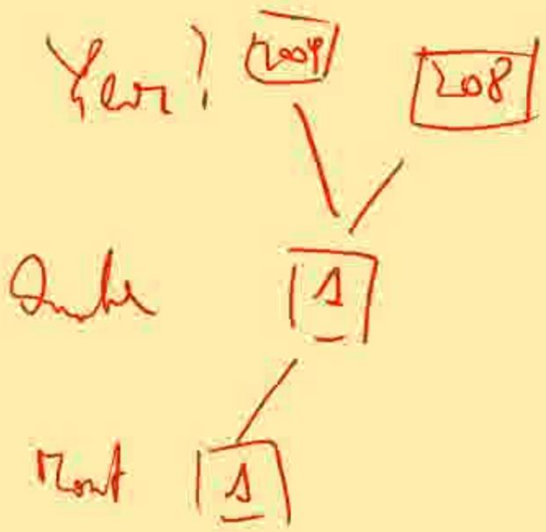
DIMENSIONAL HIERARCHIES

Date **Month → Quarter → Year**

PkDate	Month	Quarter	Year
20080101	1	1	2008
20080102	1	1	2008
...			
20090101	1	1	2009
20090102	1	1	2009

DIMENSIONAL HIERARCHIES

UNIVERSITÀ DI PISA



Date **Month → Quarter → Year**

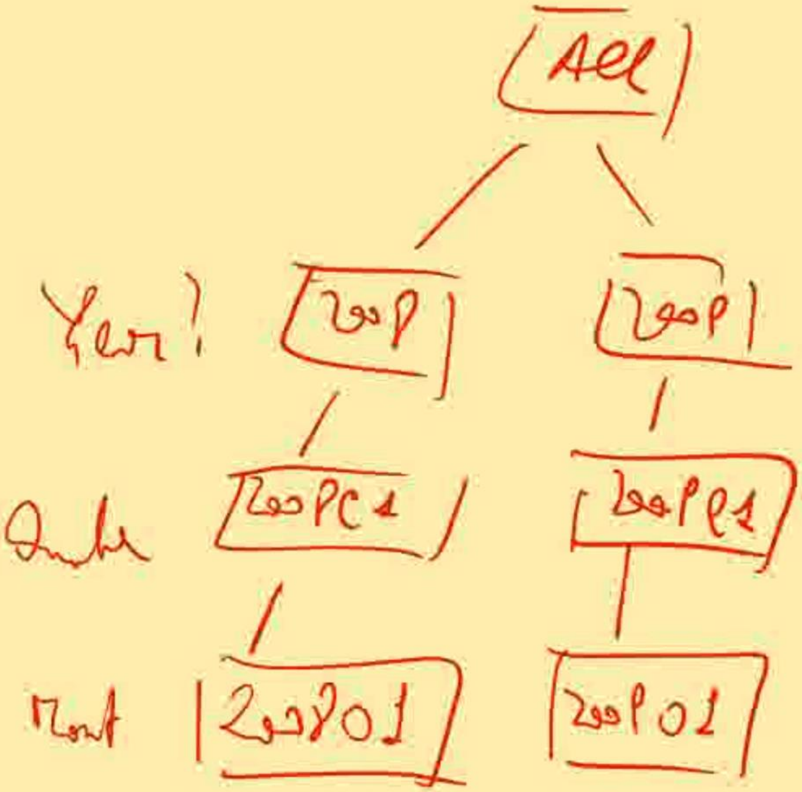
PkDate	Month	Quarter	Year
20080101	1	1	2008
20080102	1	1	2008
...			
20090101	1	1	2009
20090102	1	1	2009

DIMENSIONAL HIERARCHIES

Date **Month → Quarter → Year**

PkDate	Month	Quarter	Year
20080101	200801	2008Q1	2008
20080102	200801	2008Q1	2008
...			
20090101	200901	2009Q1	2009
20090102	200901	2009Q1	2009

DIMENSIONAL HIERARCHIES



Date **Month → Quarter → Year**

PkDate	Month	Quarter	Year
20080101	200801	2008Q1	2008
20080102	200801	2008Q1	2008
...			
20090101	200901	2009Q1	2009
20090102	200901	2009Q1	2009

In a hierarchy we want for each child a unique parent, this means we can **uniquely** associate to a **fact** the chain of **aggregations** at different levels of detail

DATA MODELS FOR DW

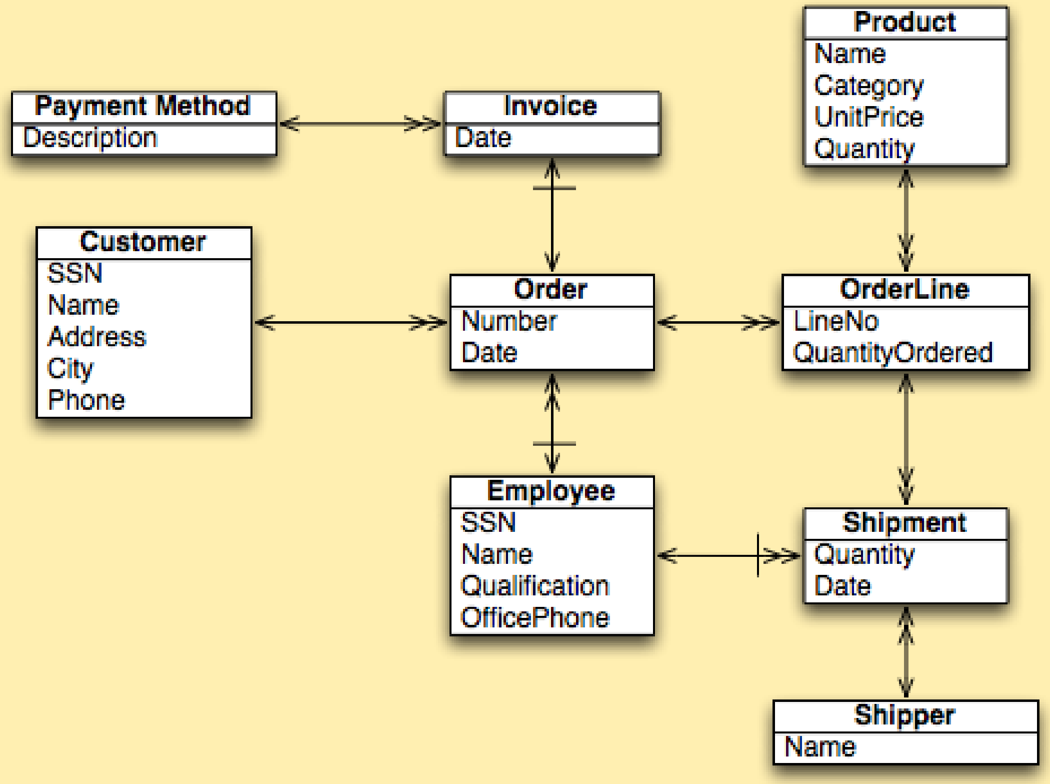
To define the structure of a DW the following formalism are used, called **data models**:

The **Dimensional Fact Model** (DFM) is a graphical conceptual model used to analyze problems, given user requirements.

The **Relational Data Model**, as a logical model to design a solution

The **Multidimensional Model** (called **Cube**), useful to understand OLAP operations.

GOAL: AN ORDER DATA MART



DATA BASE

Number of product ordered,
by product, **by** customer, **by** month

Total revenue **by** product category,
by customer, **by** year

Total revenue by customers of Italy **by**
customer city, **by** year, **by** quarter

BUSINESS QUESTIONS

Basics of a formalism to model

facts,

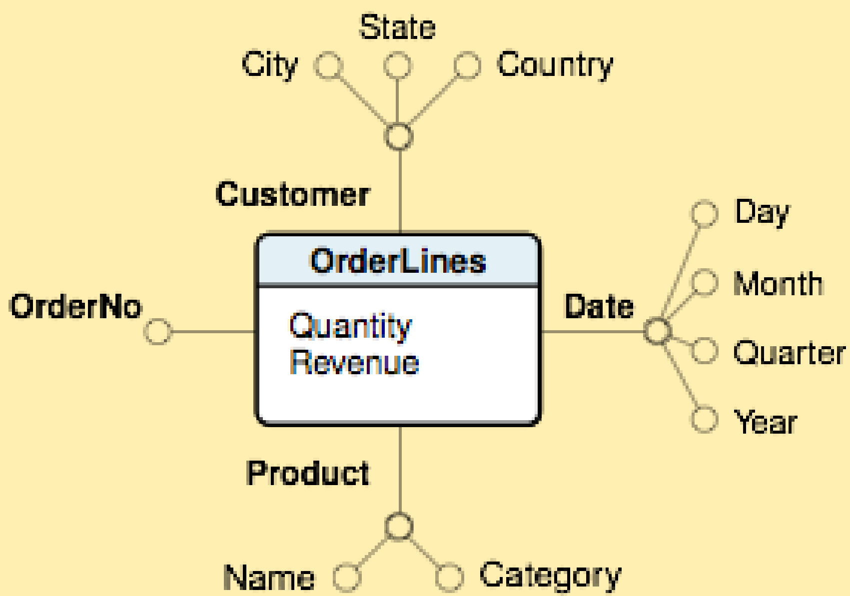
measures,

dimensions,

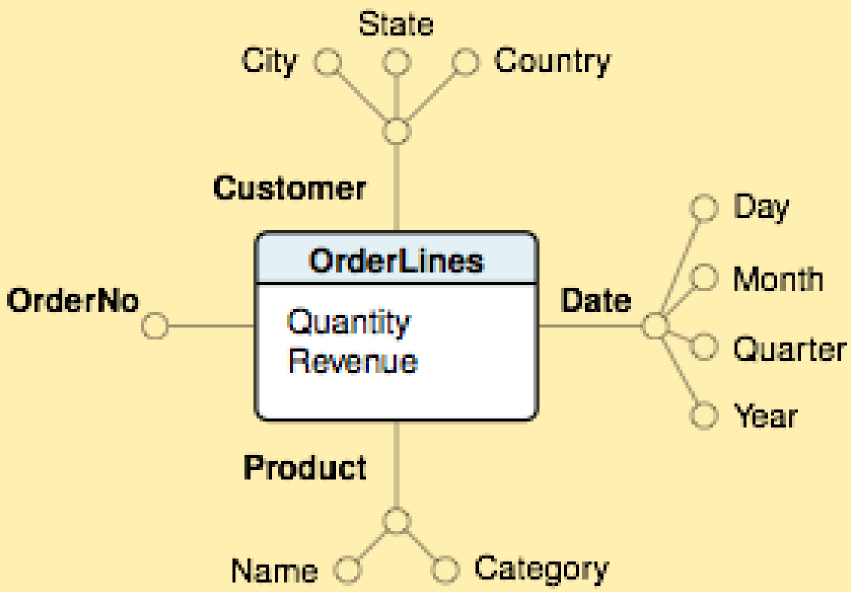
dimensional attributes.

A dimension without **attributes** is called **degenerate**

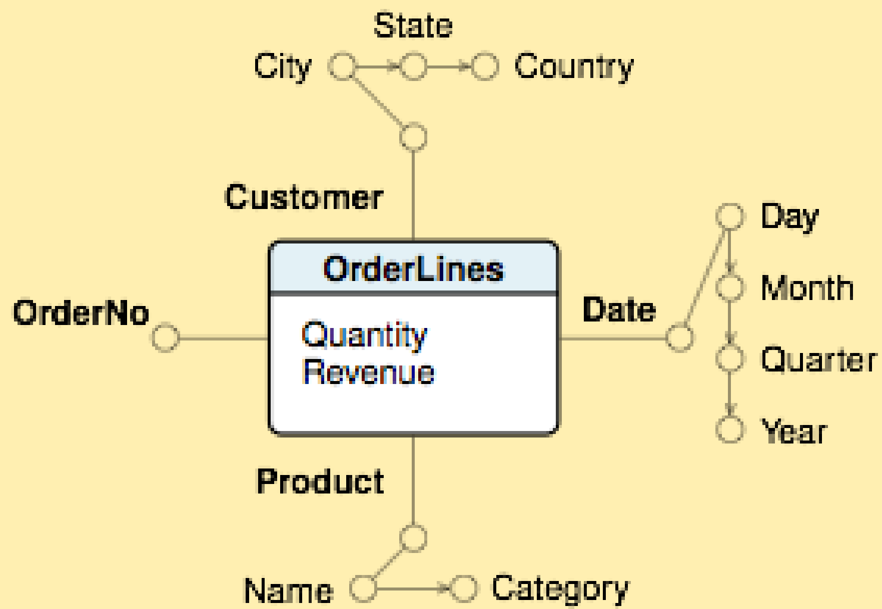
Later on other formalism features and **how to model**...



A DATA MODEL FOR CONCEPTUAL DESIGN: DIMENSIONAL ATTRIBUTES WITH HIERACHIES



Without hierarchies



With hierarchies

Let us assume that a key business process of interest has been identified together with a sample of analysis to perform to support decisions. The primary job is to understand the requirements.

Let us assume that we have understood the requirements and we want design a data mart.

INITIAL CONCEPTUAL DESIGN OF A DATA MART

Università di Pisa

Step 0: Requirements gathering.

Requirements gathering focuses on the study of business processes and on analysis relevant for decision making.

A not useful requirement analysis (a business question to answer):

Why is my business not meeting the targets?

A useful business question:

Total revenue, **by** customer, **by** product

Aggregation (metric) **Measure** **Dimensions**

Alternative: A report example

Step 1: Identify the Granularity of the Fact

The first fundamental decision to be taken is the meaning of the fact.

What is the **grain** ?

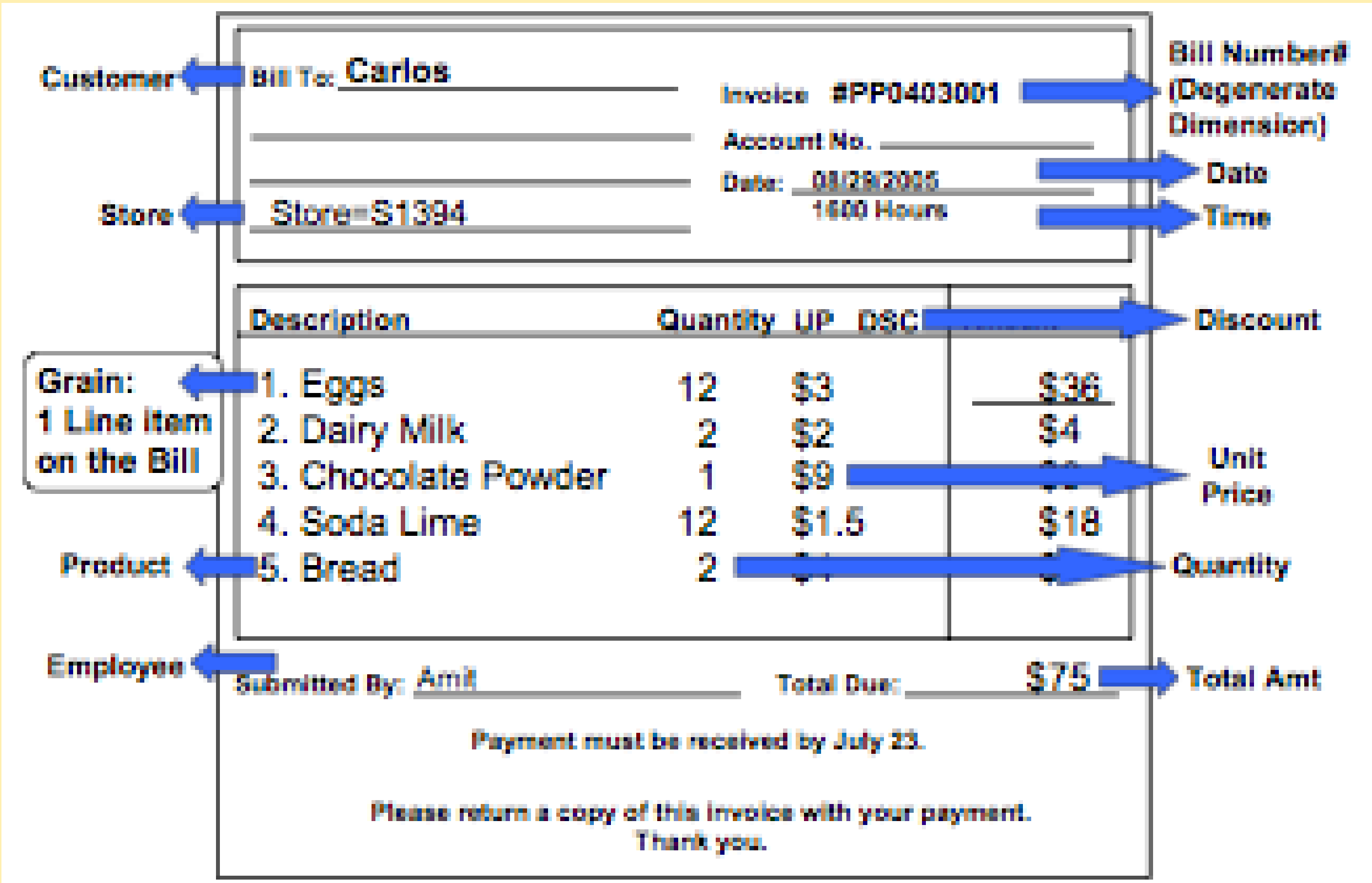
Identifying the grain also means deciding on the level of detail you want to be made available in the dimensional model. The more detail there is, the lower the level of granularity.

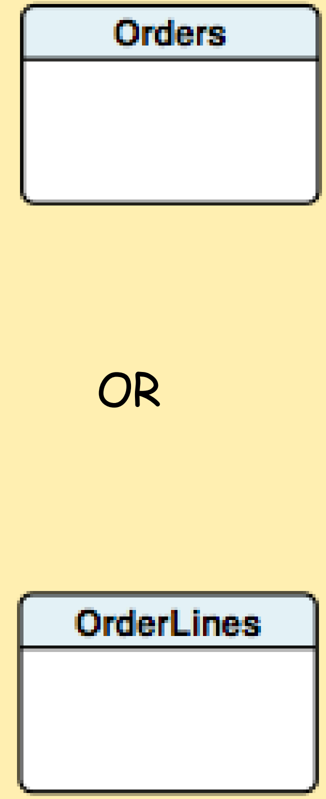
Remember:

1. **Grain is the precision with which the measurements are taken.**
2. **Grain determines measures and dimensions and dimensions determine grain !**

Example: Analyses are about customer orders. **What is an Order**?

UNIVERSITÀ DI PISA





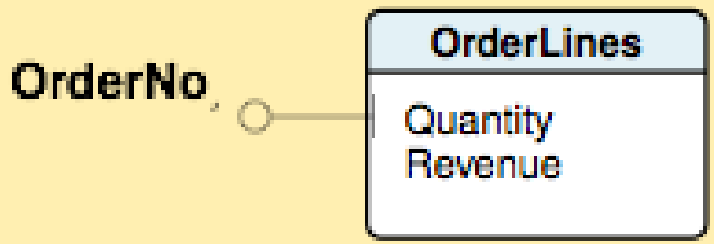
AN ORDER EXAMPLE

Step 2: Identify the Fact Measures

The **measures** of interest are **numeric values** that make sense to add.

Not everything that is numeric is a measure!

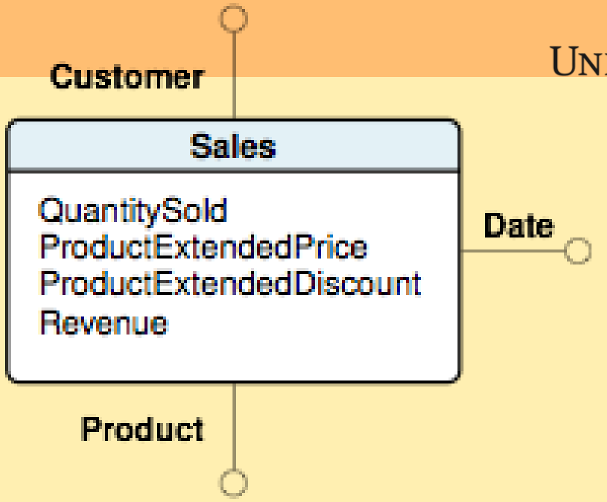
Remember:
A measure is an observation of the performance of a business process



It is important to specify a measure Type.

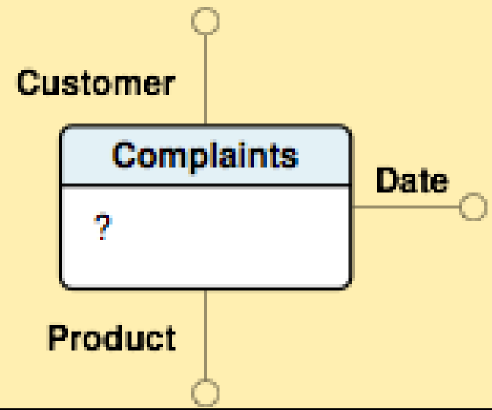
MEASURE TYPES

Numeric (calculated) **additive**.



The measures may be missing !

Factless (better **Measureless**)



Numeric **non-additive**.

Gross Margin = Margin/Revenue ?

Unit Price ?

Step 3: Identify the Fact Dimensions

Identify the **dimensions to** give fact measures their **context.**

The **Five Ws and one H questions, or the Six Ws** (?)

(from Wikipedia) are questions whose answers are considered basic in information-gathering. They are often mentioned in journalism, research, and police investigation. They constitute a formula for getting the complete story on a subject. According to the principle of the **Six Ws**, a report can only be considered complete if it answers the following questions:

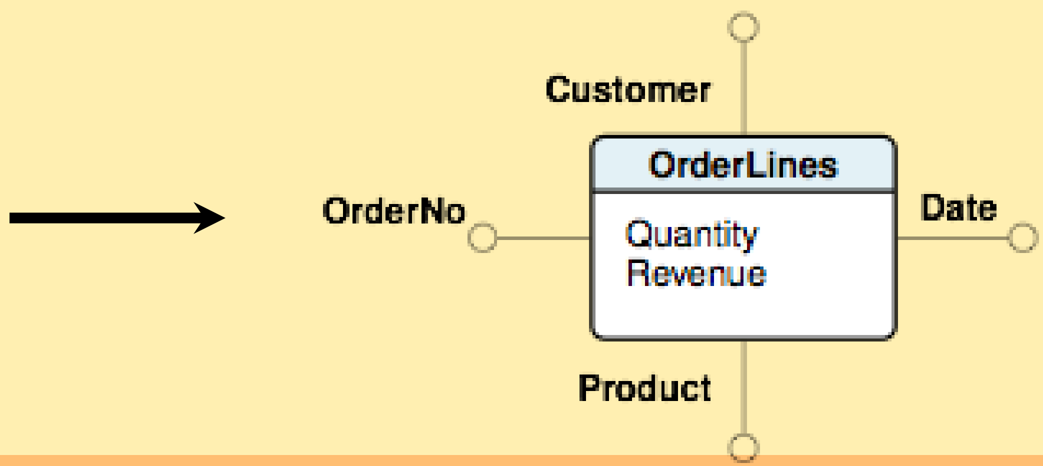
Who is it about? **What** happened? **When** did it take place?

Where did it take place? **Why** did it happen? **How** did it happen?

Step 3: Identify the Fact Dimensions

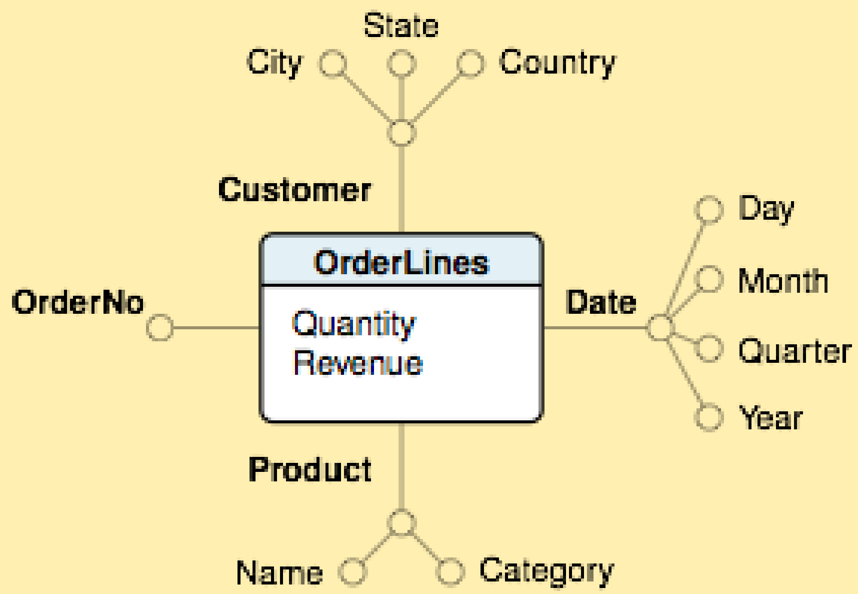
Identify the **dimensions to** give fact measures their **context.**

The **Six Ws** questions aim to identify the **variables determining the measures** and possible **intervention levers**.



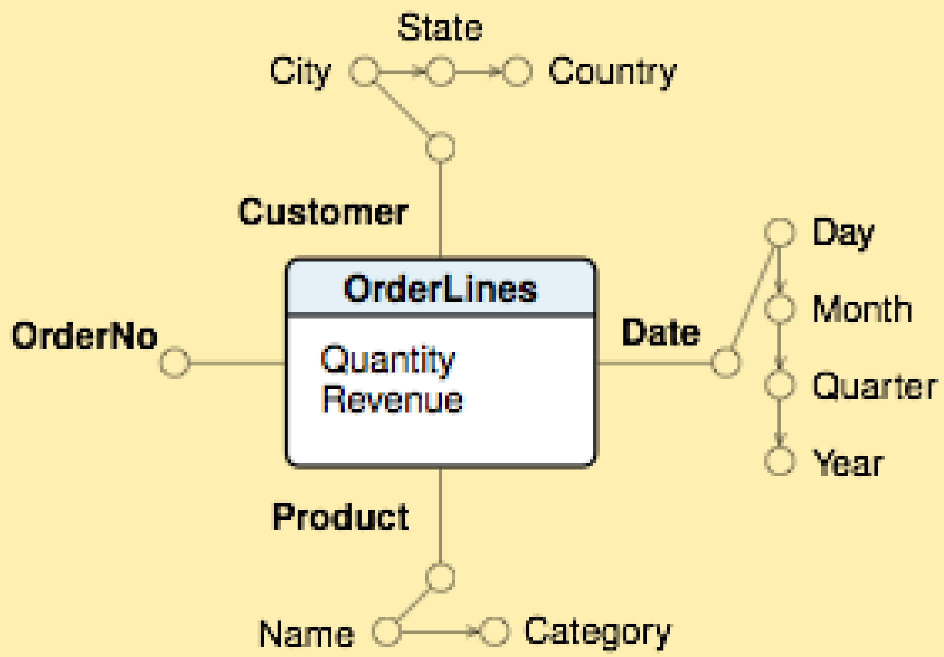
Step 4: Identify Dimensional Attributes

The dimensional attributes are important for analysis and for reports.



Step 5: Identify the Dimensional Attribute Hierarchies

Attribute hierarchies is a natural way to support interactive exploration of facts. Users understand them intuitively, because they are used to look at a summarized report and then to decide to look at a more detailed one.



CASE STUDY: University Exams

A university plans to build a data mart that would help them in analyzing the exams performance of the students in **master degree programs** in different academic sessions.

Courses have a code, which is unique, a name, whether it is mandatory or not, the teacher and department name, the credits and the semester in which a course is offered.

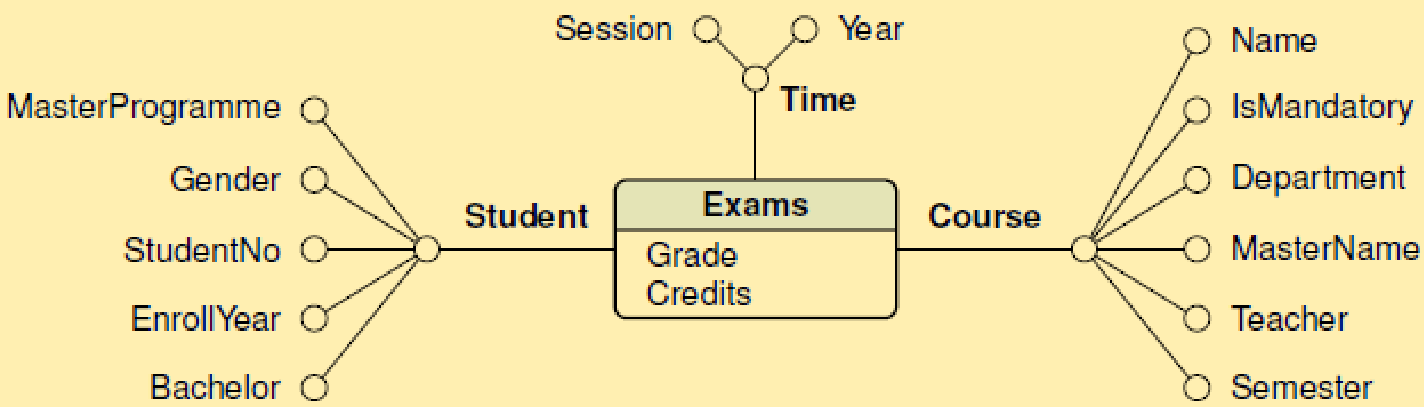
Students have a number, which is unique, the gender, the university name that awarded the bachelor degree, the name of the master degree program, the year of enrollment.

Exams have a grade, a value between 1 and 31, considered passed if the grade is greater than 17, the exam session, the academic year. Failed exams are registered too.

1. Number of exams passed, and number of exams failed, **by** course name, **by** academic year, and **by** session.
2. Number of exams failed, by the course name, by academic year, **by** session, and **by** bachelor university name.
3. For a specified master degree program and student's enrollment year, the average grade of passed exams and the total number of credits given, **by** student gender.
4. For the current academic year, average exams grade, number and the percentage of students who passed the exam, **by** the course name, and **by** session.
5. For a specified master degree program and courses with a number of exams passed of less than 3, the number of exams, **by** the course name, **by** academic year.

CASE STUDY: University Exams

	Fact granularity
Description	A fact is the occurrence of an exam
Preliminary dimensions	Student, Time, Course
Preliminary measures	Grade, Credits

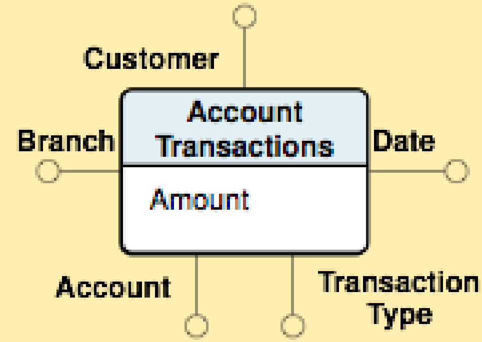


FACTS TYPES

Transaction

One fact per transaction (an event that occurs at a specific point in time)

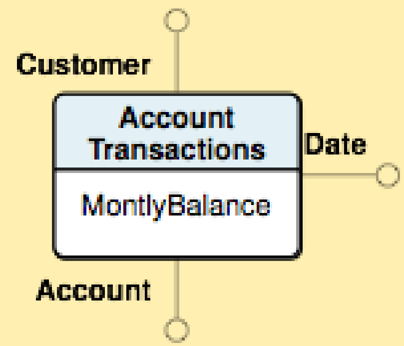
Example: A transaction for an individual account of a customer of a bank



Periodic

One fact for a group of transactions made over a period of time.

Example: The **amount** is the **monthly balance** for all transaction against an individual account of a customer of a bank.

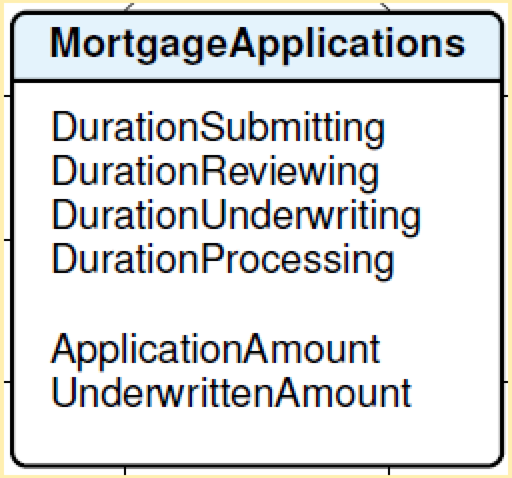


FACTS TYPES (cont.)

UNIVERSITÀ DI PISA

Accumulating One fact for the entire lifetime of an evolving event that has a duration and change over time

Example: The life time of a mortgage application.



Facts and measures only...

MORE ABOUT DATA MART CONCEPTUAL MODELLING

Degenerate dimensions

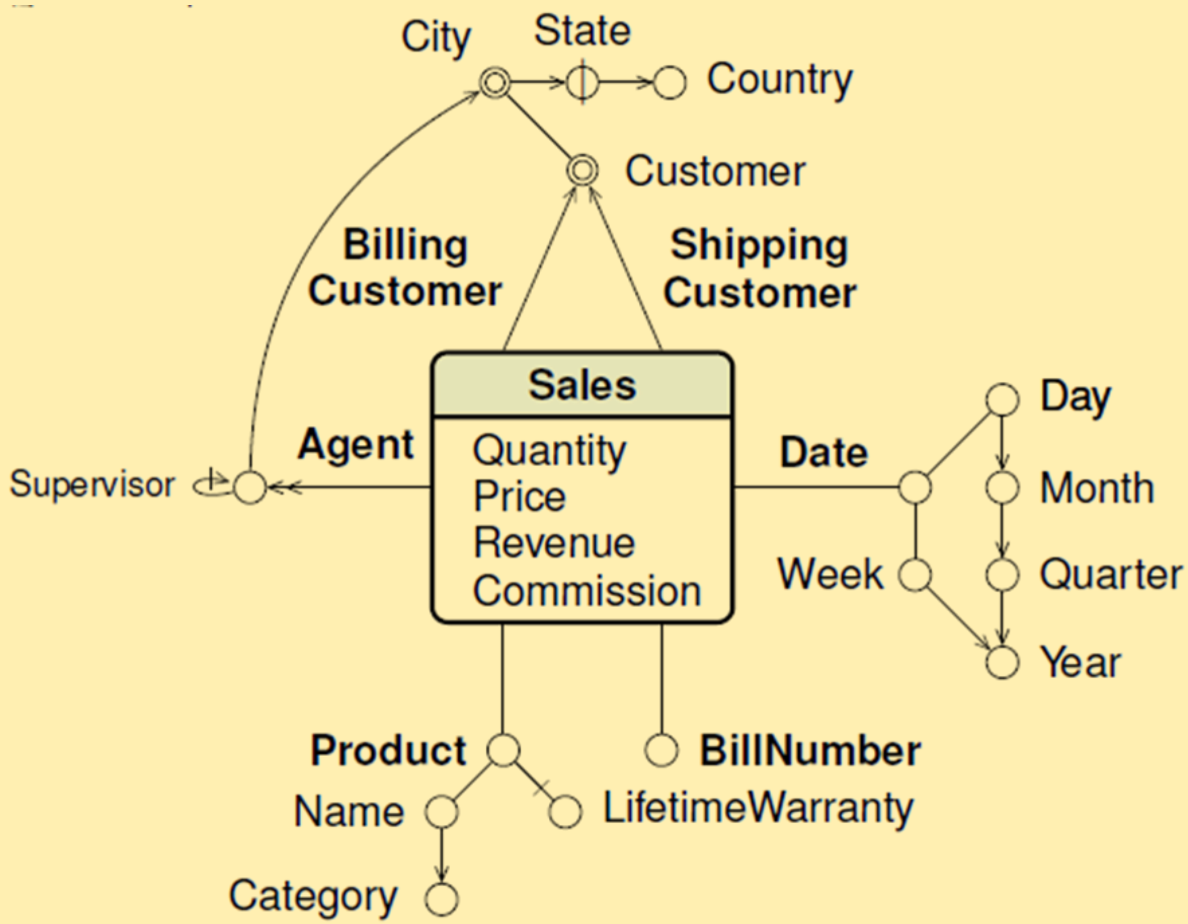
Facts descriptive attributes

Optional dimensions or attributes

Multivalued dimensions

Hierarchies types

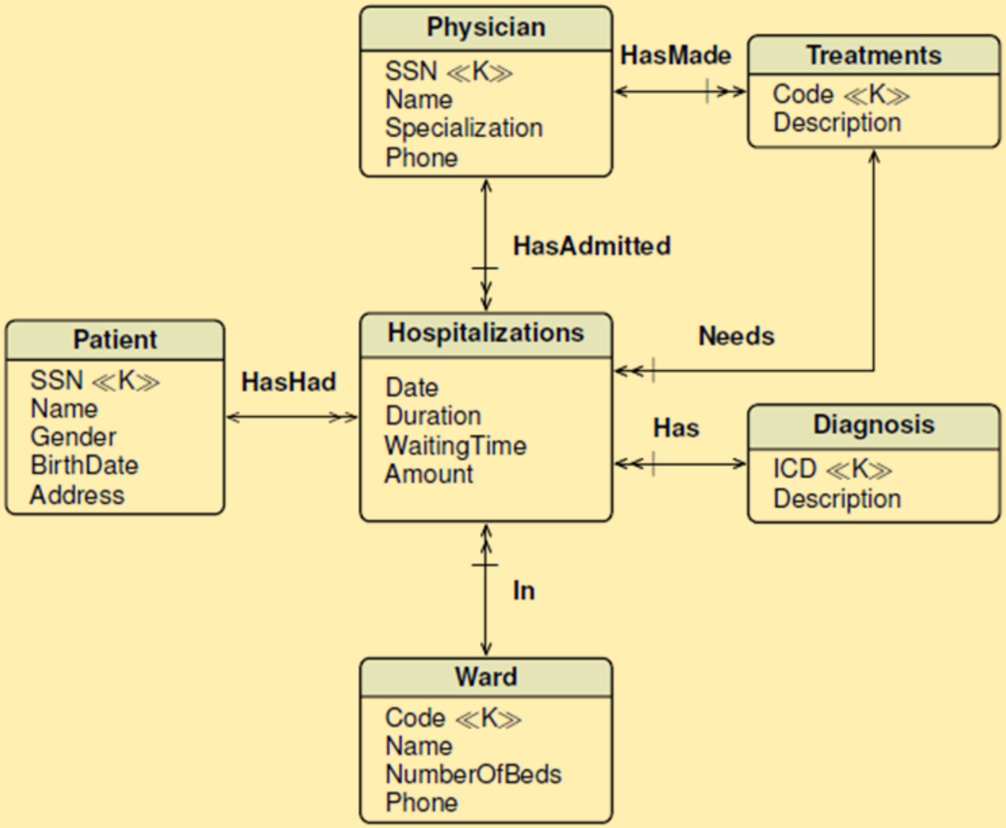
Shared hierarchies



LAB: HOSPITAL

UNIVERSITÀ DI PISA

An hospital needs a DM to extract information from their operational database with information about inpatients treatments.



1. Total billed amount for hospitalizations, **by** diagnosis code and description, **by** month (year).

2. Total number of hospitalizations and billed amount, **by** ward, **by** patient gender (age at date of admission, city, region).

3. Total billed amount, average length of stay and average waiting time, **by** diagnosis code and description, **by** name (specialization) of the physician who has admitted the patient.

4. Total billed amount, and average waiting time of admission, **by** patient age (region), **by** treatment code (description).

OPEN LAB: AIRLINE COMPANIES

Requirements analysis
Number of unoccupied seats in a given year, by flight code, by company name (or type), by class, by departure time (time, day, month, year)
Number of unoccupied seats in a given class and year, by flight code, by company name, by class, by departure (destination) city (country, continent).
Number of unoccupied seats and revenue of the Alitalia company, by year, by month, by destination country.

Hospitalization			
Requirements analysis	Dimensions	Measures	Metrics

Fact granularity	
Description	
Preliminary dimensions	
Preliminary measures	

Data Mart Conceptual Schema