Data warehousing

Mario Guarracino

Introduction

- Since the mid-nineties, it became clear that the databases for analysis and business intelligence need to be separate from operational.

- In this lecture we will review the characteristics of the data warehouses and data mart, analyzing the differences from the operational systems.

- We will also analyze the functional aspects of the data warehouses, giving some details concerning the implementation aspects.

Definition of data warehouse

Data storage

+

Procedures to acquire, organize and elaborate

=

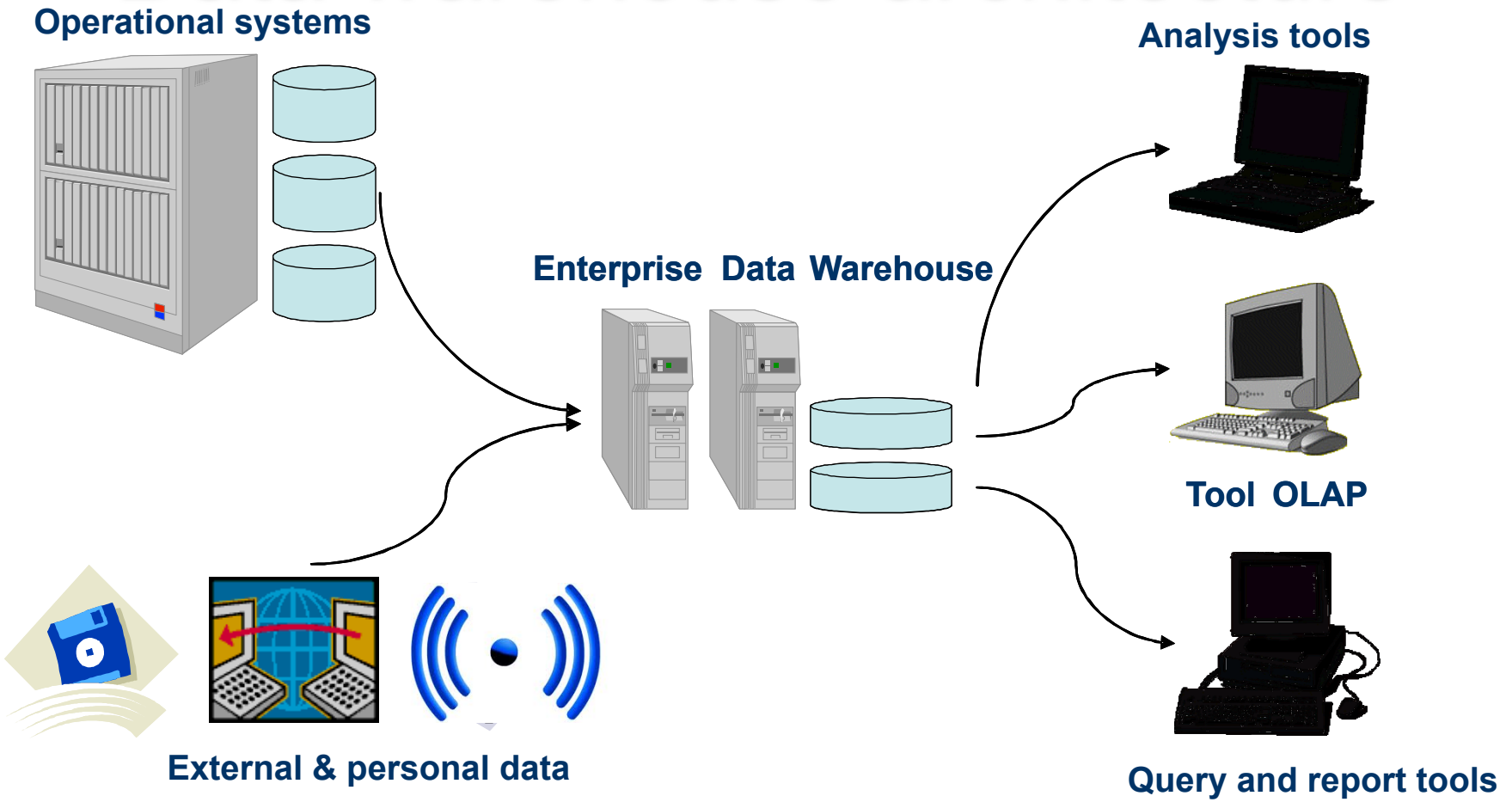
Data Warehouse

- Data warehousing = the set of activities to design, implement and use a data warehouse.

Motivation

- A number of reasons lead us to realize a data warehouse separated from OLTP database:
  - **Integration**: data warehouses require data from different sources
  - **Quality**: the data affect the results
  - **Efficiency:** analysis must be rapid
  - **Extension of time**: the data must have sufficient historical depth.

Data warehouse architecture



Characteristics

- Collection of data with the following properties:
  - **Oriented to subjects**: consider the data of interest for managers, and for organizational processes.
  - **Integrated**: across the enterprise and departmental.
  - **Historicized**: with long horizon
  - **Consolidated** (aggregate): do not care "who" but "how many".
  - **Denormalized**: redundancies allow faster response times.
  - **Offline**: data updated periodically.

Data mart

- Departmental data warehouses collect data from a specific company.
- Specialized system that brings together the data needed for a department.
- Implemented by creating specific views to applications.
  - A data mart contains marketing information to customers and sales transactions, results of campaigns, ...
- Materialized views subsets with departmental focus on certain subjects.

ETL tools

- Extract, Transform, Load: set of tools to extract, transform and load data
- In the first stage, the data are extracted from operational databases.
  - Initial pull, incremental
- In the transformation phase inconsistencies, duplications, and not admissible values are eliminated.
- The corrected data are transformed and loaded into the data warehouse.

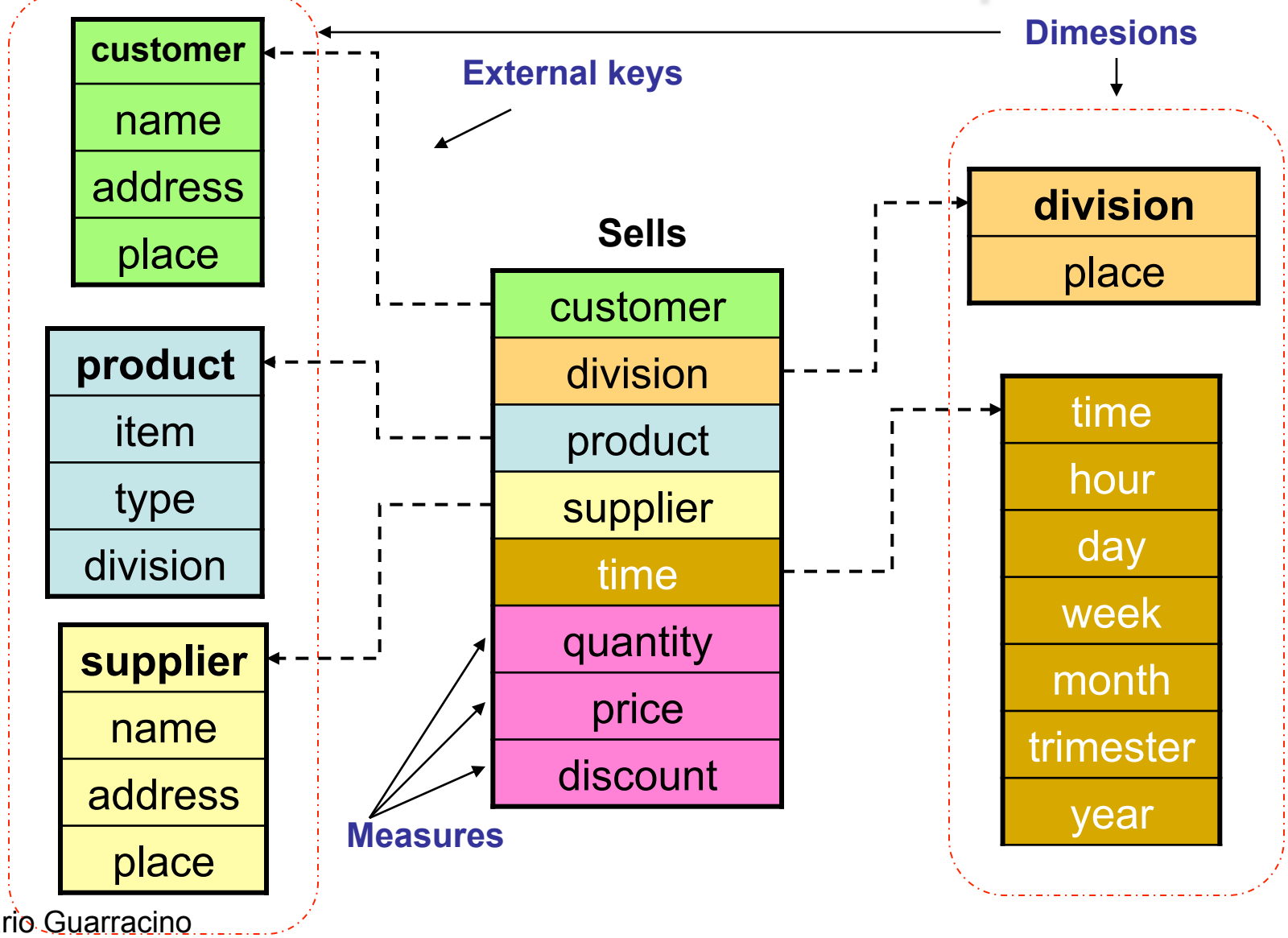
Multimensional cube analysis

- The conceptual modeling of a data warehouse using:
- **Star Schema**: A single object (the fact table) in the middle connected to a number of objects (dimension table).
- **Snowflake schema**: A refinement of star schema where the dimensional hierarchy is represented explicitly (by normalizing the dimension tables).
- **Galaxies**: Multiple fact tables share dimension tables.

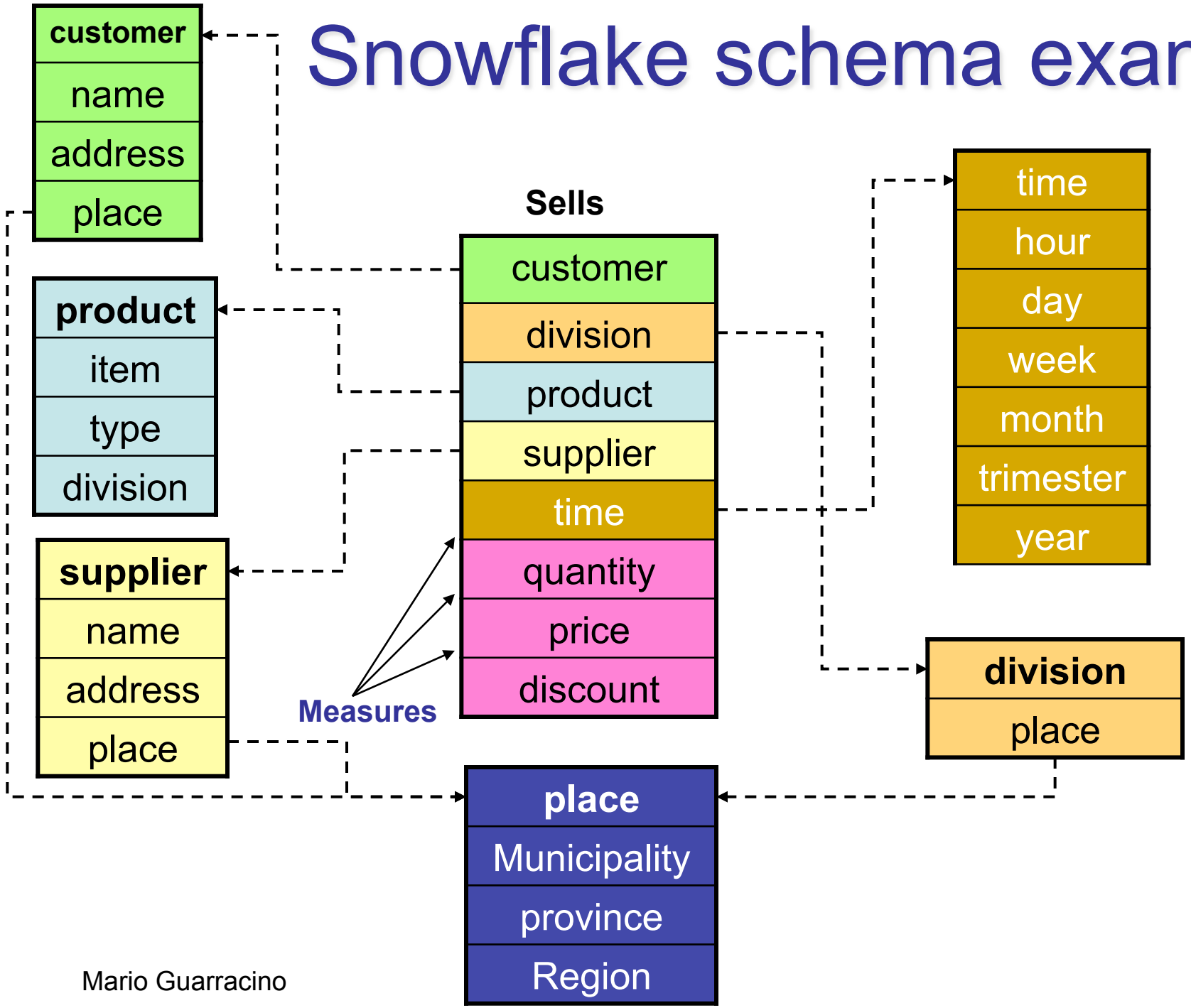
Star schema

- A **fact** is an event of interest to the enterprise.
  - sales, shipments, purchases, ...
- The **measures** are attributes that quantitatively describe the fact from different points of view.
  - number of units sold, unit price, discount, ...
- One **dimension** determines the minimum granularity of representation of the facts.
  - the product, point of sale, the date
- A **hierarchy** determines how instances of a fact can be aggregated and select - describes one dimension.

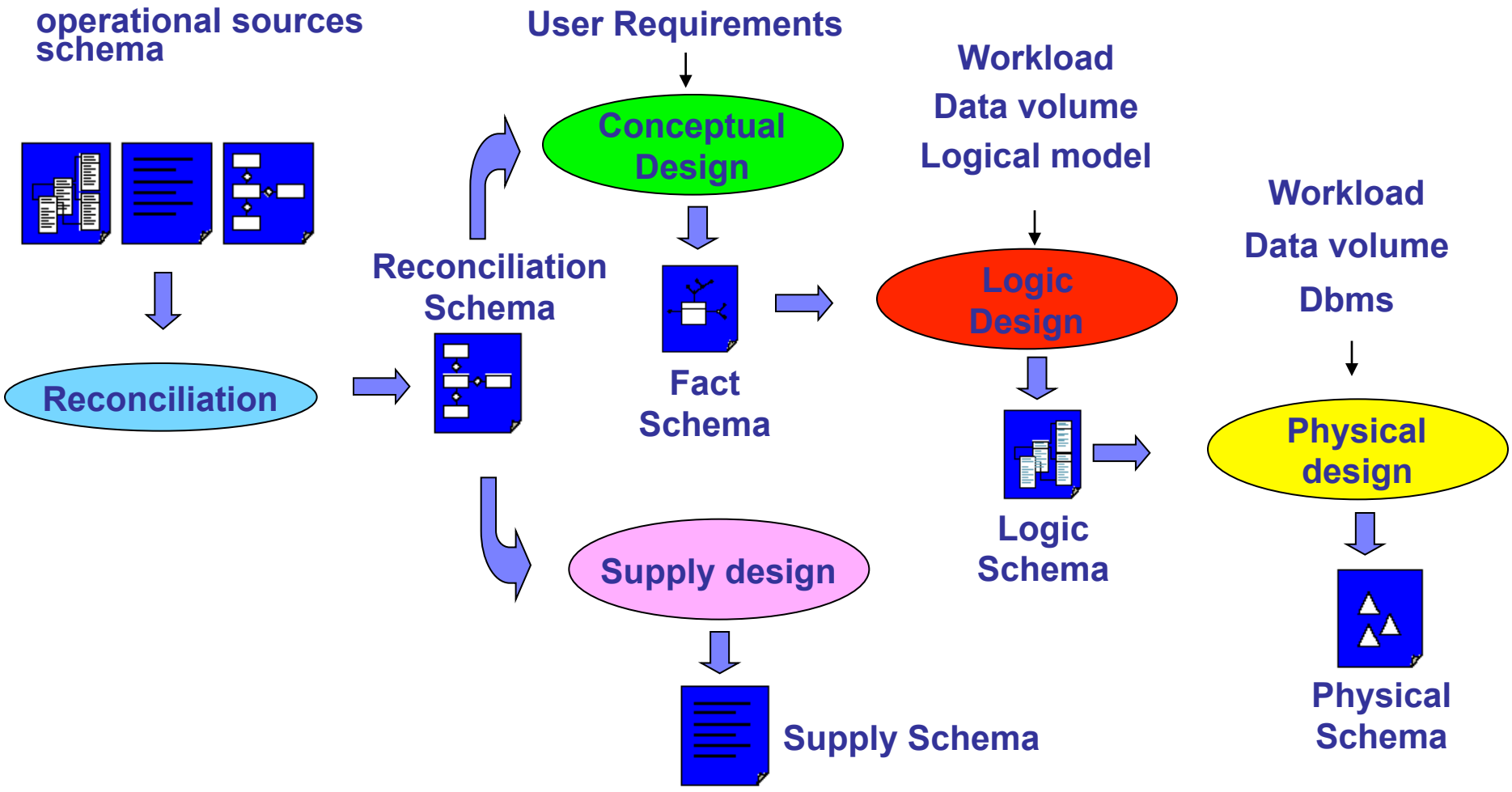
Star schema example



Snowflake schema example



Data mart design

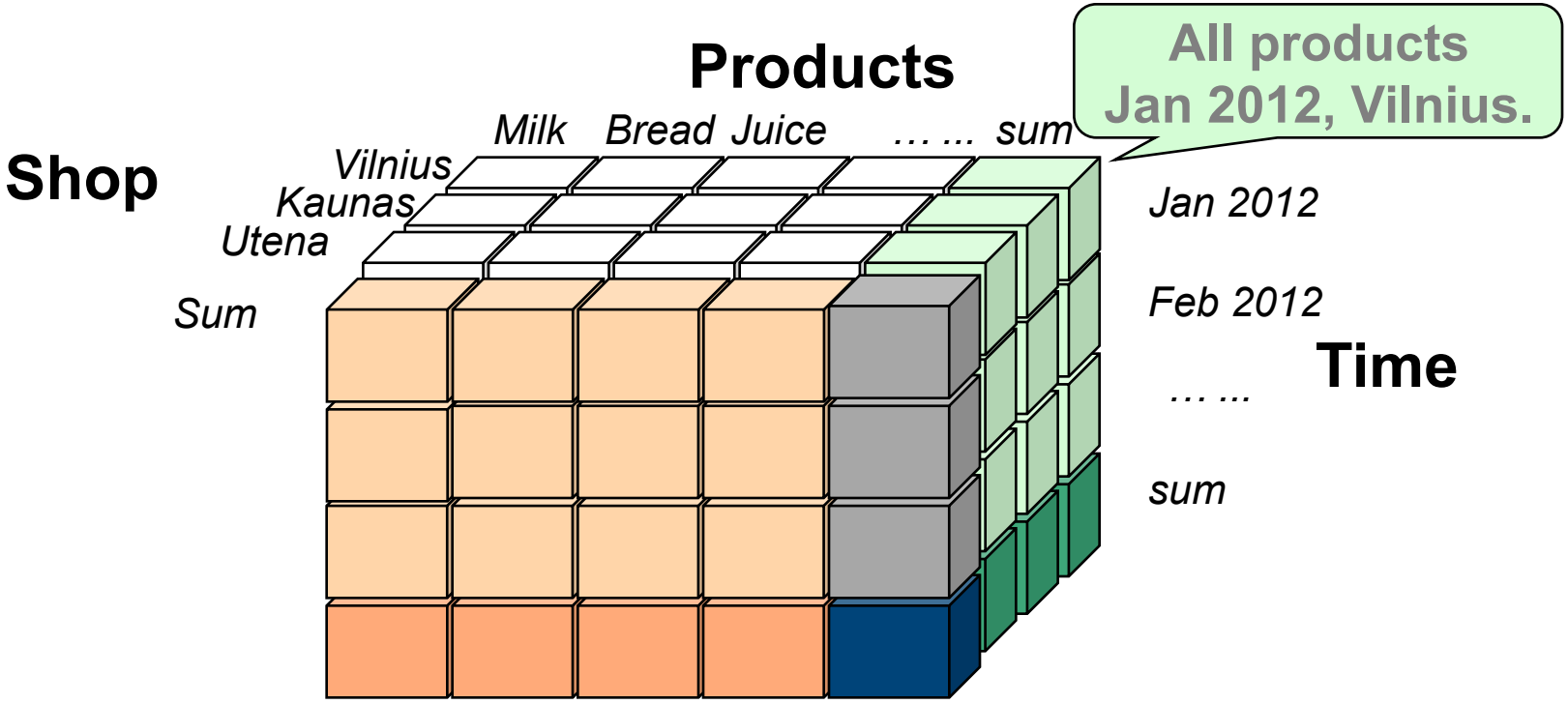


Data cubes

- A fact table connected to a n dimension tables can be represented by an n-dimensional data cube.
- Each dimension contains a hierarchy of values and a cube cell contains the aggregated values
  - count, sum, max, ...
- They represent a natural evolution of spreadsheets.
- In Excel, they are called pivot tables.

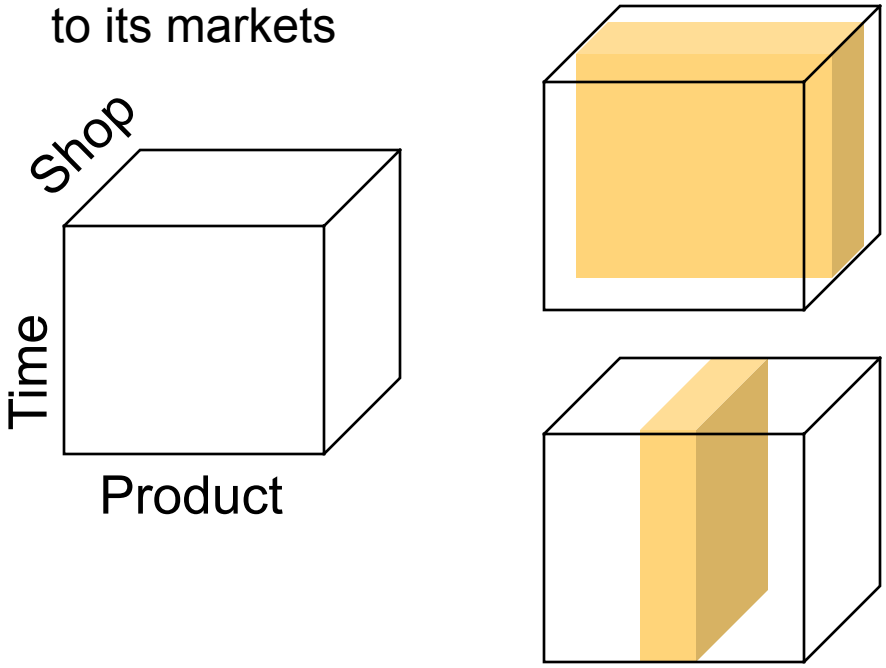
Example

- **Fact table**: Sells
- **Dimensions**: {time, product, shop}
- **Measure**: number of sold units

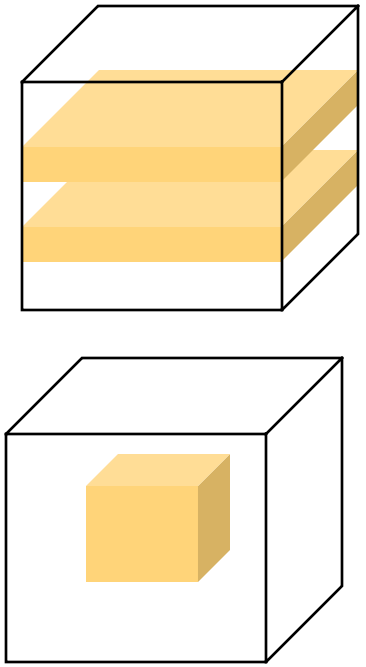


Example

Regional Manager examines
the sale of products
in all periods, with respect to
to its markets



Financial manager examines the
sale of products in all markets,
comparing current and previous
period



Product manager examines
the sale of a product
in all period and in all markets

The strategic manager focuses on
a category of products, regional area
and a average horizon time

Operations on cubes

- **Roll up**: summarize data the total volume of sales by product category and region
- **Roll down, drill down, drill through**: switches from low level to a high level of detail for a particular product, find detailed sales for each salesperson and each date
- **Slice and dice**: select & project
  - Sales of beverages in the south area in the last 6 months
- **Pivot**: reorganization of the cube

Data quality

- The business intelligence analysis depends heavily on the quality of the input data.
- The extracted data from various sources and collected data marts may have abnormalities that are identified and corrected.
- We will see what are the main techniques to identify and remove anomalies and simple solutions to improve the accuracy and efficiency of the learning algorithms.

Validation

In a dataset, data can be affected by:

- Incompleteness
  - **Missing values**
  - Records without the valuesof certain attributes
  - Valuesavailable only in aggregate form
  - Noise
- Measurement errors
  - Record and **outliers**
  - Duplicate data
  - Inconsistency
  - Contradictions between values or between records

Missing data

- **Missing completely at random (MCAR)**: A variable is missing completely at random if the probability of a miss is the same for all units.
  - Example: if each survey respondent decides whether to answer a question by rolling a die and refusing to answer if a “6” shows up.
- If missing data are MCAR, then throwing out cases with missing data does not bias your inferences.

Missing data

- **Missing at random (MAR)**: the probability a variable is missing depends only on available information.
  - Example: if sex, race, education, and age are recorded for all the people in a survey, then "earnings" is missing at random if the probability of nonresponse depends only on these other, fully recorded variables.
- It is often reasonable to model this process as a logistic regression, where the outcome variable equals 1 for observed cases and 0 for missing.

Missing data imputation

- Possible solutions:
  - Ignore the tuple.
  - Enter the value manually.
  - Insert a global constant (e.g. 0, ∞, ⊥).
  - Enter the average value of the attribute.
  - Enter the average value of the attribute for the class of tuples to which it belongs.
  - Enter the amount most likely, calculated by a suitable method (e., e. Regression).

Outlier

- Definition by Hawkins [Hawkins 1980]:
  - *"An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism".*
- Statistics-based intuition:
  - Normal data objects follow a "generating mechanism", e.g. some given statistical process.
  - Abnormal objects deviate from this generating mechanism.
- Detecting measurement errors:
  - Sensors data may contain measurement errors.
  - Removing such errors can be important for data mining and data analysis tasks
  - "One person's noise could be another person's signal."

http://www.dbs.ifi.lmu.de/~zimek/publications/KDD2010/kdd10-outlier-tutorial.pdf

Sample variance

- ***Sample variance***: given a set of m values $(x_1, x_2, \ldots, x_m)$ with mean μ:

$$\bar{\sigma}^2 = \frac{1}{m-1}\sum_{i=1}^{m} s_i^2 = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{\mu})^2$$

- A smaller sample variance implies a lower dispersion of values around the mean.
- To return the measure of dispersion to the original scale of the observations, we use the ***sample standard deviation***:

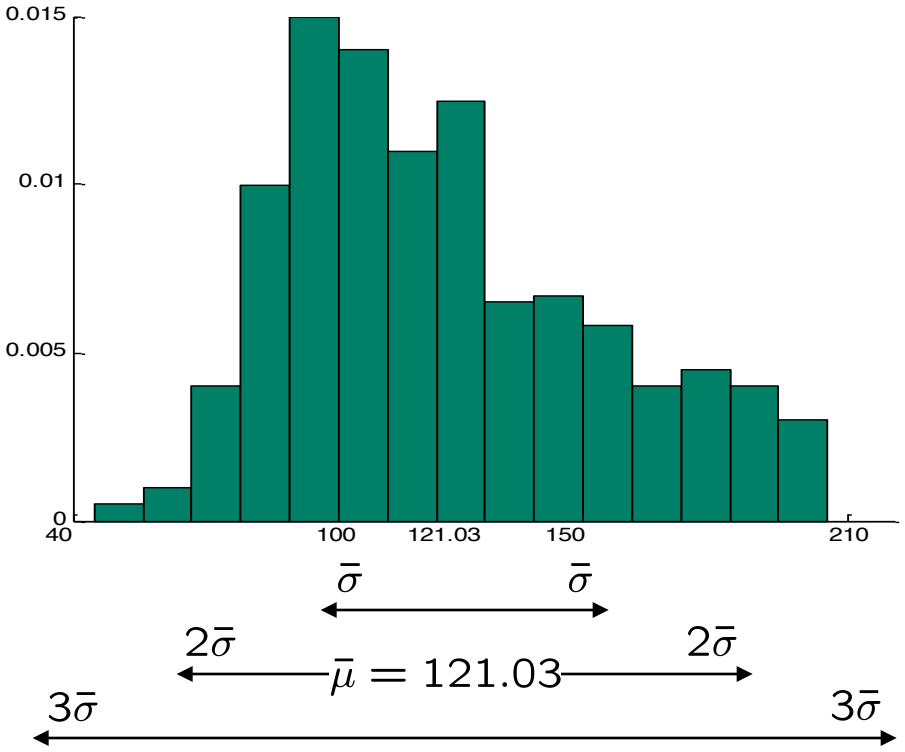
$$\bar{\sigma} = \sqrt{\bar{\sigma}^2}$$

Sample variance

- The variance can be used to delimit the interval around the sample mean where it is reasonable to expect that the sample values fall (interval without outliers).
- Normal distribution
  - The range $(\mu \pm \sigma)$ contains ~68% of the observed values
  - The range $(\mu \pm 2\sigma)$ contains ~95% of the observed values
  - The range $(\mu \pm 3\sigma)$ contains ~100% of the observed values
- Arbitrary distribution
  - Even if the distribution is significantly different from normal, it is possible to derive intervals in which samples are expected fall.

Tchebysheff theorem

Given a number γ and a set of m values $a = (x_1, x_2, \ldots, x_m)$, a percentage at least equal to $(1\text{-}1/\gamma^2)$ of values falls within the range $(\mu \pm \gamma\sigma)$ i.e. at no more than γ standard deviations from the sample mean.



Data streams variance

- When data is produced by sensors, they are not available all at the same time.
- If we analyze a stream of data in windows, each with n_k elements of mean μ_k and variance $\sigma^2{}_k$, then:

$$n_{i,i+1} = n_i + n_{i+1}$$

$$\mu_{i,i+1} = \frac{n_i \mu_i + n_{i+1} \mu_{i+1}}{n_{i,i+1}}$$

$$\sigma^2_{i,i+1} = \frac{n_i \left(\sigma_i^2 + \mu_i^2 \right) + n_{i+1} \left(\sigma_{i+1}^2 + \mu_{i+1}^2 \right)}{n_{i,i+1}} - \mu^2_{i,i+1}$$

http://infolab.usc.edu/csci599/Fall2003/Data%20Streams/Maintaining%20variance%20and%20k-medians%20over%20data%20stream%20windows.pdf

Summary

- We have seen:
  - Data warehouses & data mart;
  - Fact tables, dimensions and indices;
  - Star schemas and snowflake;
  - Multidimensional analysis;
  - Missing data & outliers

Exercise

- Consider data from ese1.txt available from the course webpage. Suppose data arrive at a speed of 10 sample/ sec and max window is 60 secs.

- Write a program in your preferred compiled/interpreted language to enhance data quality and find outliers using Tchebysheff theorem.

- Submit your results to mario.guarracino@cnr.it by Friday, December 2nd, 2012, with subject

  BIfloT: Outliers detection

Exercise

- Starting from the problem you identified:
  1. Specify the input data (sensors, eternal,…).
  2. Identify facts, dimensions and measures.
  3. Build up the fact tables to answer one or more questions.
- Wrap up these concepts in max 5 slides and add to previous presentation.
- Submit your results to mario.guarracino@cnr.it by Friday, December 2nd, 2012 with subject

  BIfIoT: Project