Business intelligence

Unit 2 – Datawarehouse and OLAP
S2-1 – Datawarehouse

Introduction to datawarehouse

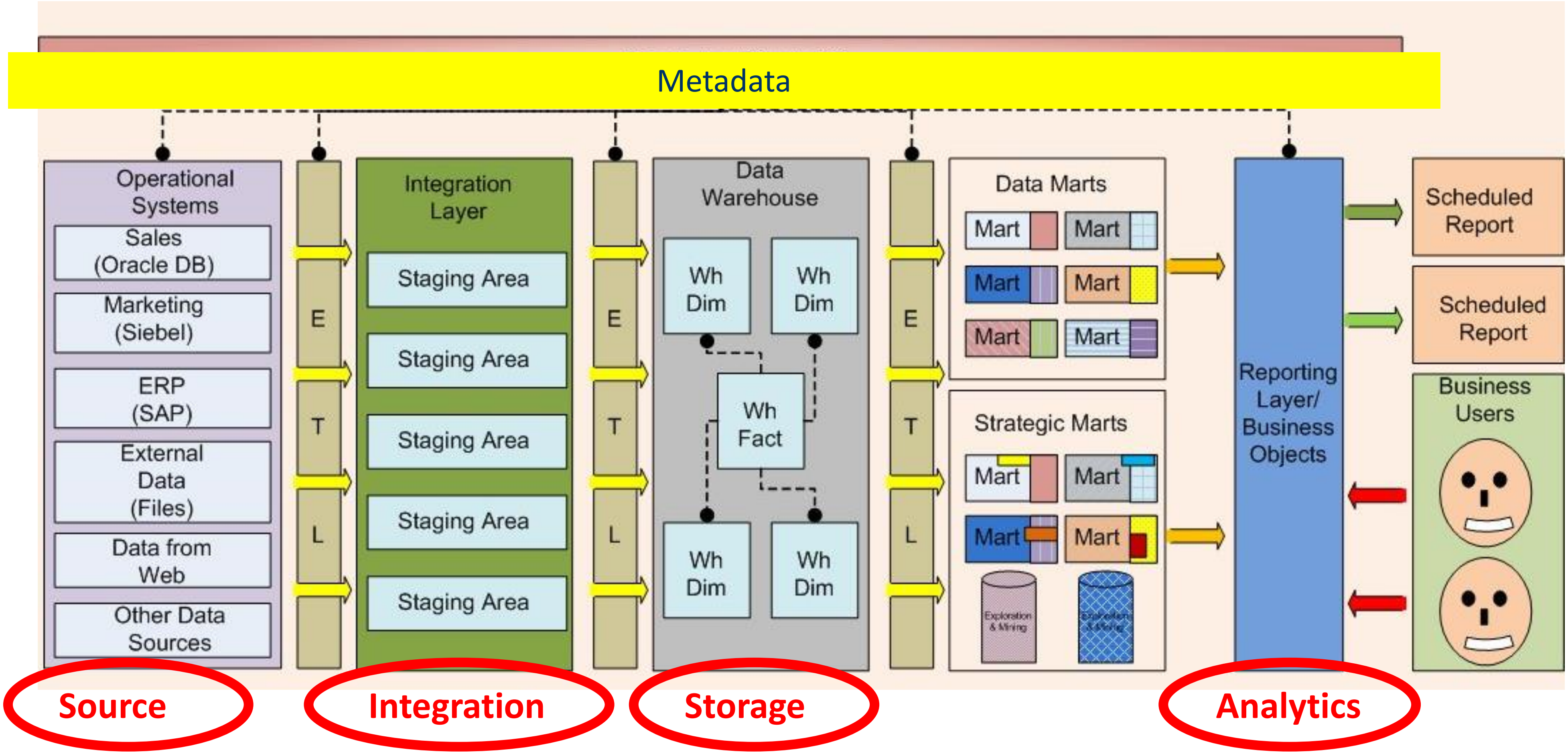
- Objective: to provide infrastructure for the DSS (Decision Support System) in an organization
  - We start from basic systems to organizational processes.
  - These systems may have several operational databases.
- DSS have new requirements
  - We want to extract knowledge from the databases and historical operational.
  - In order to:
    - Analysis of the organization
    - Make predictions
    - Define strategies

Introduction to datawarehouse

- You can still use the traditional system:
  - Maintaining daily transactional work in the original information systems (known as OLTP, On-Line Transactional Processing).
  - Basic data analysis is done in real time on the same database (known as OLAP, On-Line Analytical Processing).
- but:
  - Efficiency problems in daily work due to complex queries that are made when there is low load.
  - Efficiency problems in the analysis because there is no any specific design. Not possible in real time.

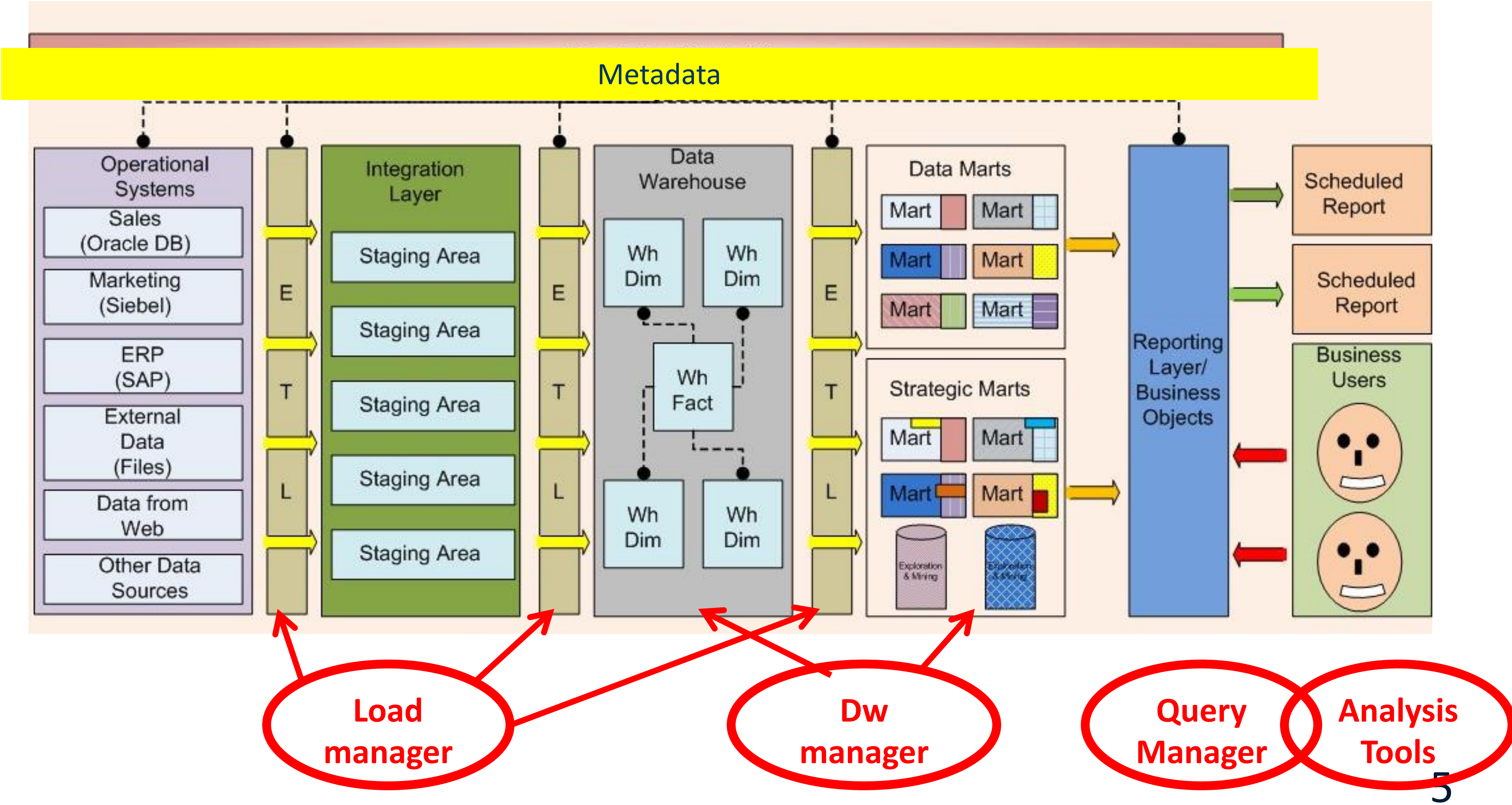
Conceptual Architecture and Components

http://solutioning.businessintelligenceconsultancy.in/



Conceptual Architecture and Components

http://solutioning.businessintelligenceconsultancy.in/

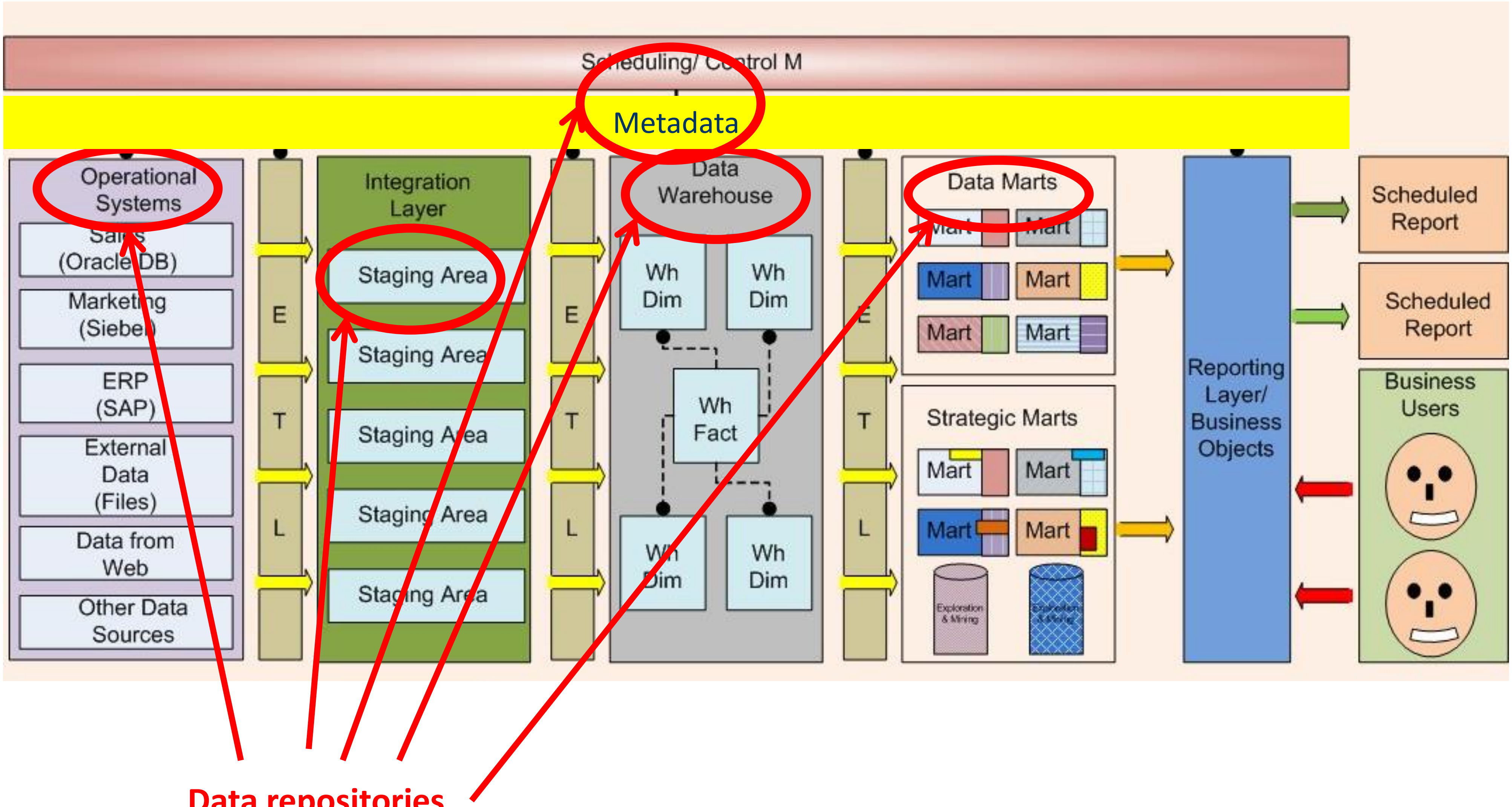


DW Components

- Load manager: runs ETL tasks
  - Extraction
  - Transformation
  - Load
- Dw manager (server): it allows to define and maintain the datawarehouse: data definition, aggregation, views, index, backup, etc..
- Query manager: Query execution, monitoring, ad-hoc forms, etc.
- Access tools: tools to design queries and reports, tools to develop end-user applications, OLAP tools, data mining tools, enterprise Information Systems (EIS)

Conceptual Architecture and Components

http://solutioning.businessintelligenceconsultancy.in/



Data repositories

Data

- Data sources: files, www, xls, databases, …
- Staging area: integrated conceptual model of information.
  - E-R, Relational model.
  - Not compulsory to have it implemented, but usually convenient.
- Datawarehouse: data collection for decision making
  - Multidimensional model
- Data mart: departamental dw
- Metadata: describes an organization in terms of its business activities, the business objects, and rules on which the business activities are performed.
  - Technical meta data needs to be mapped to the business meta data.
  - Includes documentation about data sources (origen, description, aggregation level, storage, …)

Datawarehouse: Features of a DW

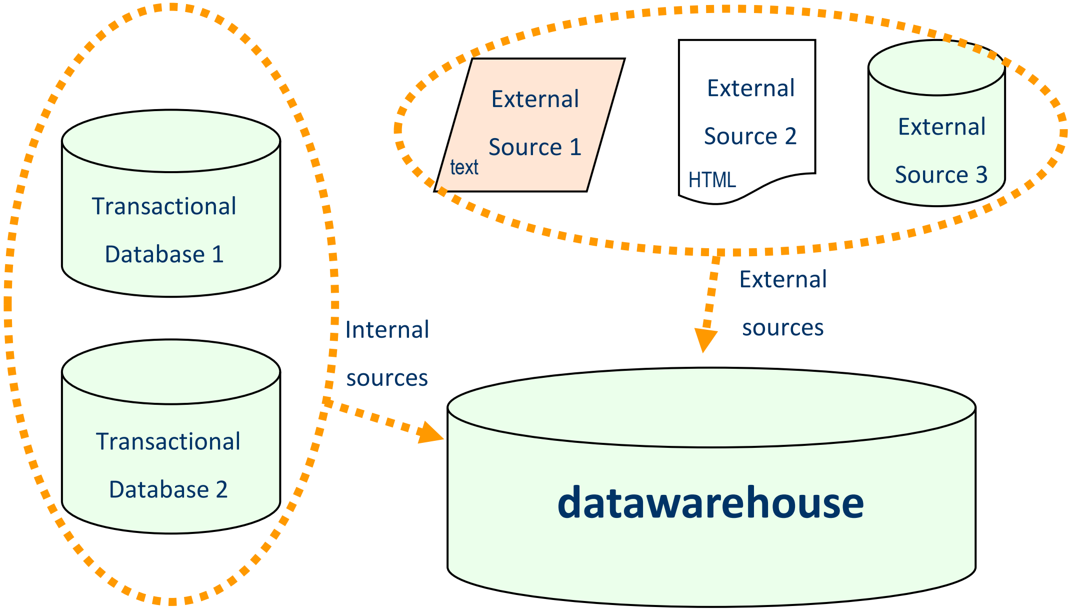
- The core of a BI architecture is the Datawarehouse
- Data Warehouse: A collection of data designed to support the decision-making processes:
  - Information Oriented (not processes)
  - integrated
  - Variable over time
  - Nonvolatile

Features of a DW

- Information oriented (not processes): dw is designed to efficiently view information on the basic activities (sales, purchasing, production, ...) of the organization, not to support the processes that take place in it (order management , billing, etc).
- Necessary information is extracted from transactional systems and efficiently stored for analysis.
- It leaves out irrelevant information.

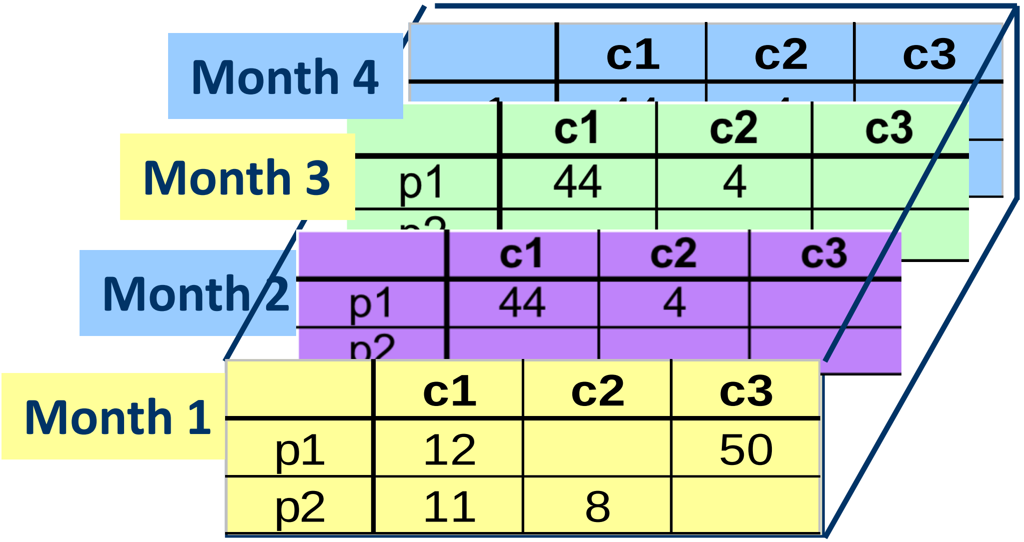
Features of a DW

- Integrated: it collects data not only from the transactional databases, but it can also include external sources



Features of a DW

- Variable in time: the data are relative to a period of time and must be periodically increased



Month 3

	c1	c2	c3
p1	44	4	

Month 2

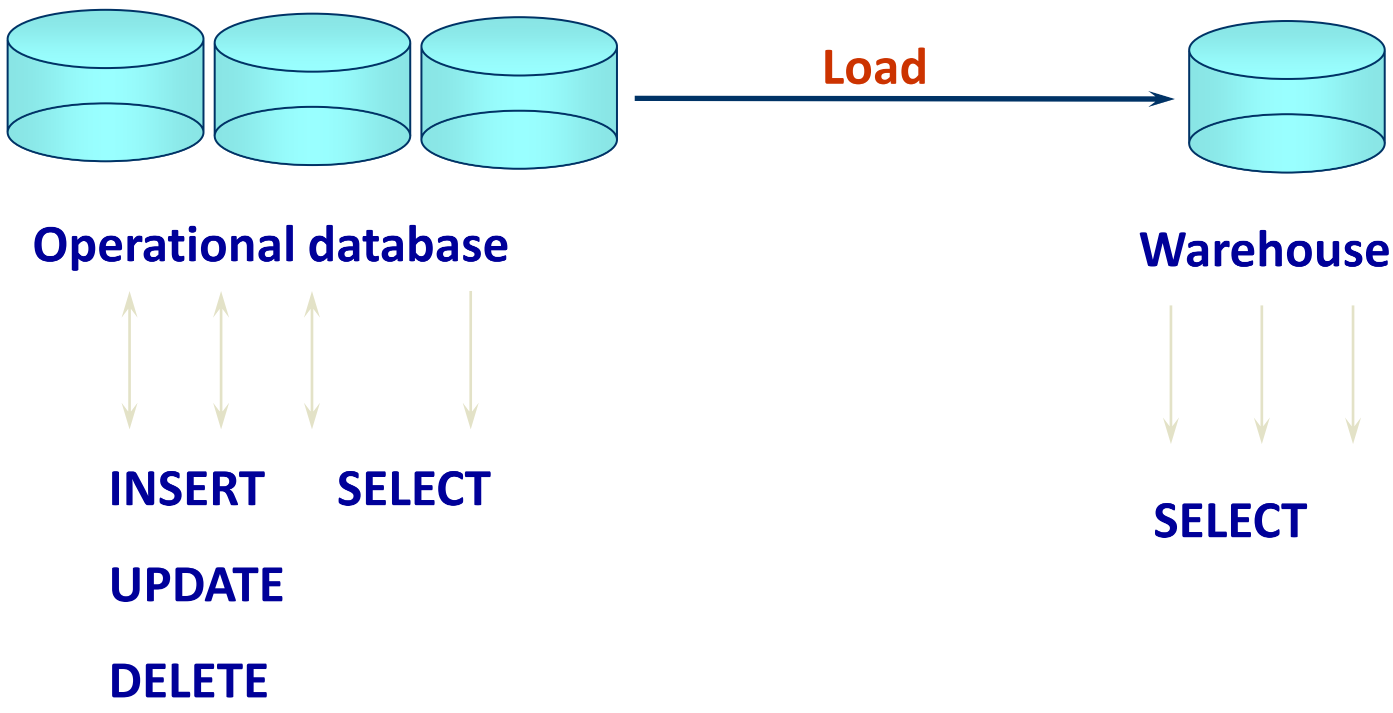
	c1	c2	c3
p1	44	4	

Month 1

	c1	c2	c3
p1	12		50
p2	11	8	

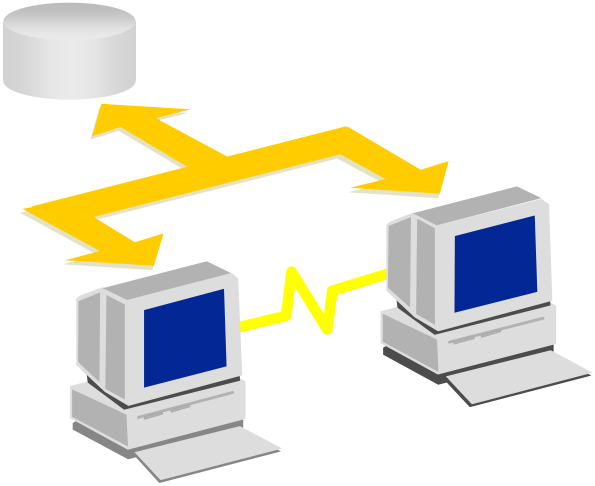
Features of a DW

- Non-volatile: the stored data are not updated, only increased

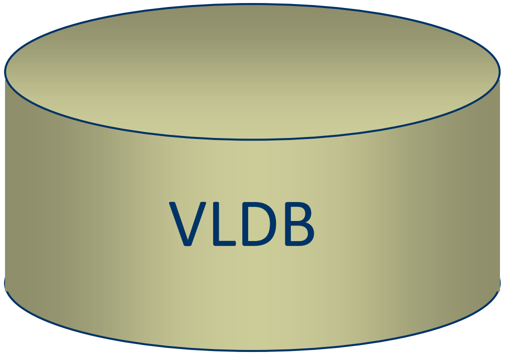


Introduction

- Advances in technology have encouraged the development of data warehouse technology
  - Big data technology
    - Parallelism
    - Hardware
    - Distributed operating systems
    - Database
    - Query languages



- VLDB
- Big memory
- Indexing techniques
- Open systems (interoperability)
- Specialized hw and sw for DW
- Tools for data analysis



OLTP vs DW

OLTP system	Datawarehouse
Stores current data	historical data (trend->include current data)
Stores detailed data	Stores summarized or aggregated data
Data are dynamic (updatable)	Data are static
Repetitive processes	Ad-hoc unforeseeable processes
Predictable usage pattern	Unpredictable usage pattern
High rate of transaction	Medium or low rate of transactions
Low response time (seconds)	Variable (usually long) response time
Directed by transactions	Directed by data analysis
Application or process-oriented	Information oriented
Supports daily decisions	Supports strategic decisions
High number of users (administrative)	Few users (managers)
Medium-size databases	Large-size databases

Advantages of using DW

- The advantages of using DW are, among others:
  - High ROI
  - Competitive advantages:
    - Information non previously available, unknown or difficult to extract and incorporate
  - Higher productivity in decision making personnel:
    - More integrated information with easy access

Problems

- Problems:
  - Understatement of resources to load
  - Complexity of integration
  - Hidden problems of source systems
  - High demand for resources
  - High cost of ownership
  - Required data are not captured
  - Increased demand from end users
  - Homogenization data
  - Data ownership
  - Long-term projects

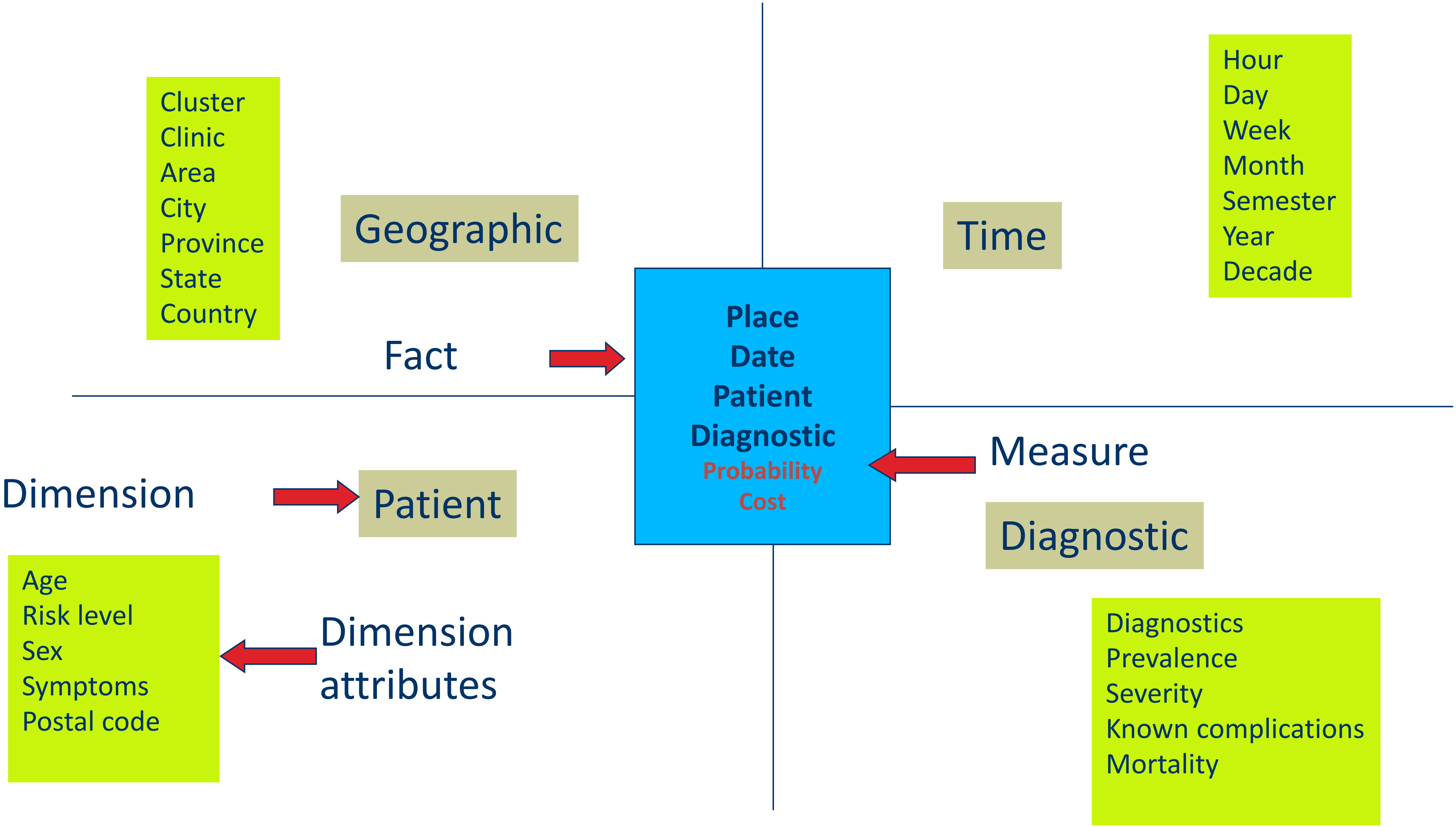
Multidimensional model

- Multidimensional model:
  - it models an activity which is subjected to analysis (fact) and dimensions that characterize the activity.
  - relevant information about the event (activity) is represented by a set of indicators (measures or fact attributes).
  - descriptive information for each dimension is represented by a set of attributes (dimension attributes).

Multidimensional model

- **Activity analized**: Diagnostics.
- Information recorded about diagnostic: “Avian influenza diagnostics has been realized in the clinic “Morales” on Oct, 11th 2012 with a probability of 85% to the patient “Joseph”.
- The geographic and temporal context are important, not the concrete diagnostic to a person.

Multidimensional model



Cluster
Clinic
Area
City
Province
State
Country

Geographic

Hour
Day
Week
Month
Semester
Year
Decade

Time

Fact

Place
Date
Patient
Diagnostic
Probability
Cost

Measure

Dimension

Patient

Diagnostic

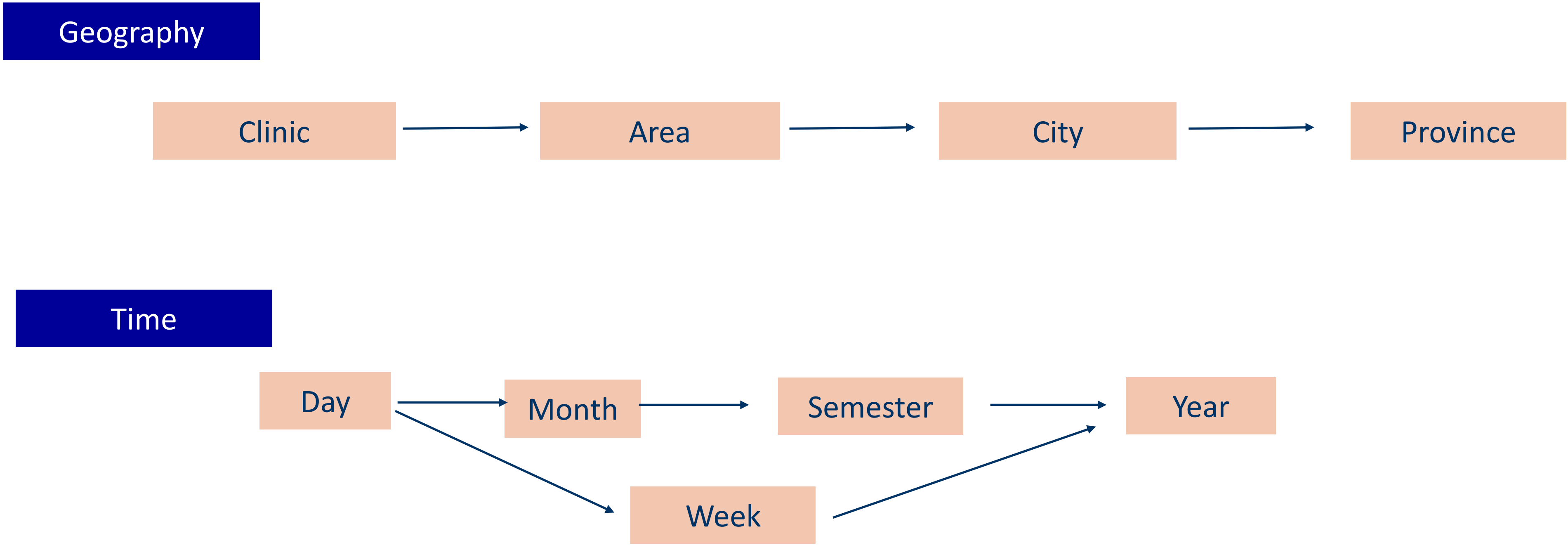
Age
Risk level
Sex
Symptoms
Postal code

Dimension attributes

Diagnostics
Prevalence
Severity
Known complications
Mortality

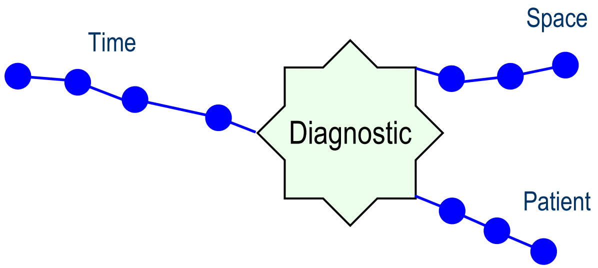
Multidimensional model

- Hierarchy: sorting between the dimension’s attributes

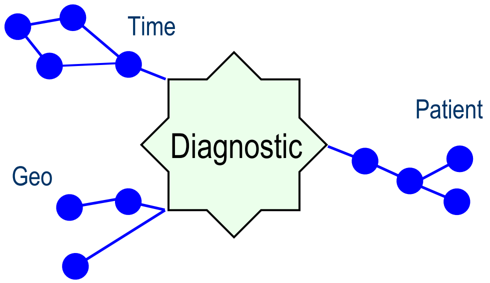


Multidimensional model

- Basic structures
- Star: lineal relation between dimension’s attributes



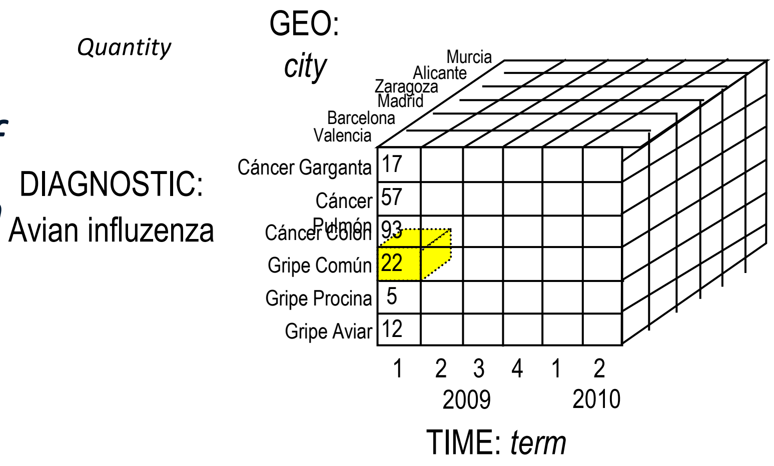
- Snowflake: non linear hierarchy

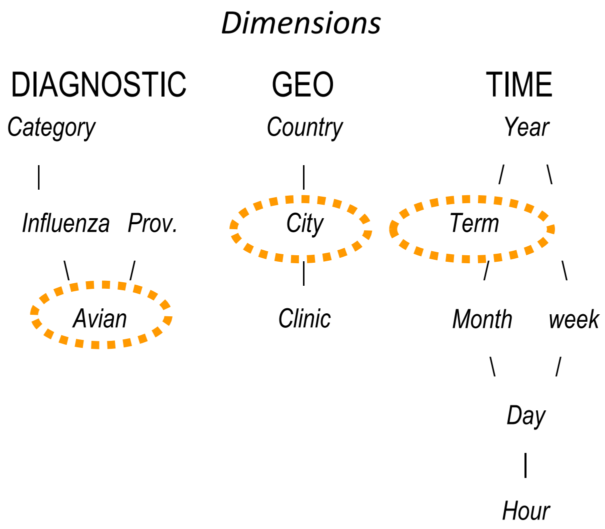


Multidimensional model

- Facts can be queried at different aggregation levels:
  - Query measures about the facts parametrized by dimensions' attributes and constrained by values of those attributes

FACT: *"The first tem of 2010 22 cases of avian influenza where diagnosed in Murcia*

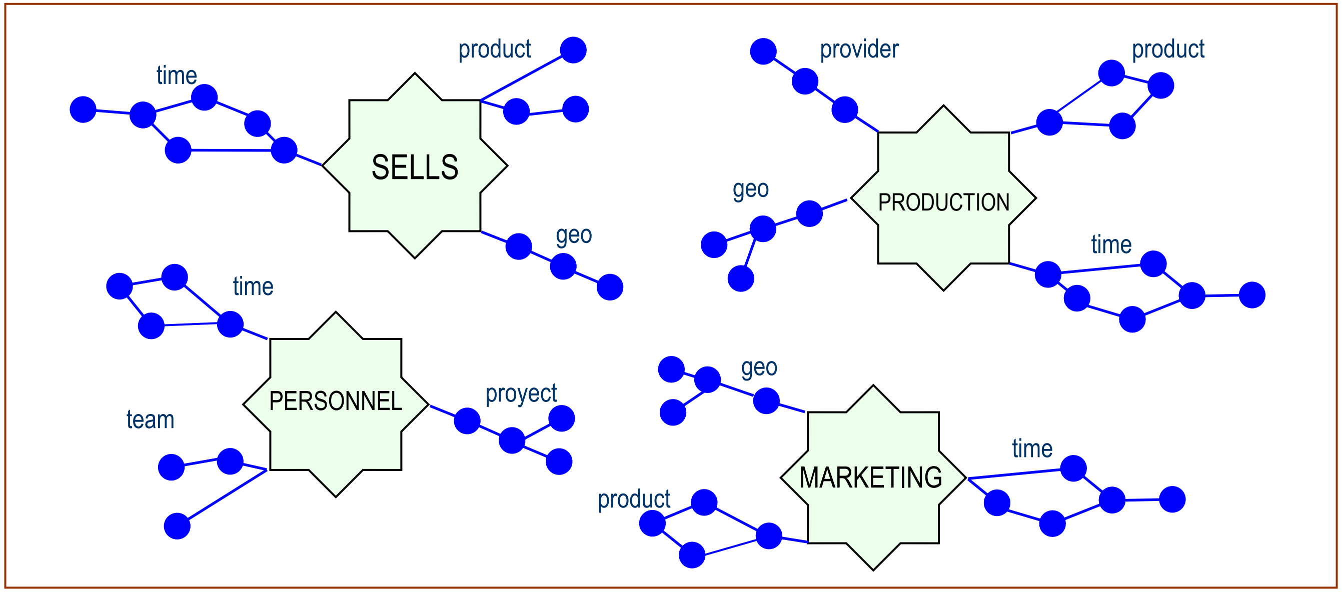




- An aggregation level for a set of dimesions is called cube.

Multidimensional model

- Not all the information can be stored in a single schema
  - Some departamental datawarehouses are needed
  - Each of these is called datamart.



Design strategies

- Datamart:
  - defined to meet the needs of a department or subdivision of the organization.
  - contains less detail and more aggregated information.
- Up-down (Inmon)
  - First define the data warehouse of the whole organization and then define data marts on him
- Bottom-up (Kimball)
  - predefine departmental data marts and then integrate them into a data warehouse for the organization

Datamart

- Datamart: subset of the datawarehouse
- The datawarehouse can be formed by several datamarts and, optionally, additional tables.
- Are defined to meet the needs of a department or subdivision of the organization.
- Contains less detailed information and more aggregate information.
- They are easier to understand and use.
- They may be the intermediate step between the data warehouse and transactional system

Datamart

- Why datamarts:
  - Provide users with access to the most commonly analyzed
    - With better response time
  - The data will already be adapted for OLAP functions or DM.
  - Give as an external source the summarized view of department users
  - Being simple, ETL processes are too.
  - Low cost.
  - Greater involvement of users in the overall data warehouse.

Data lakes

- In DW data are transformed and structured, and we keep only the data needed for the analysis
- Are we losing data?
- Data lakes store ALL source data in original format (may be not transformed)
- Easy to transform later for new DW or for data analysis
  - But schema-on-read. Is it good?
- Created for data scientist and explorers, not for business users