Advanced Data Management Technologies

Unit 2 — Basic Concepts of BI and DW

J. Gamper

Free University of Bozen-Bolzano
Faculty of Computer Science
IDSE

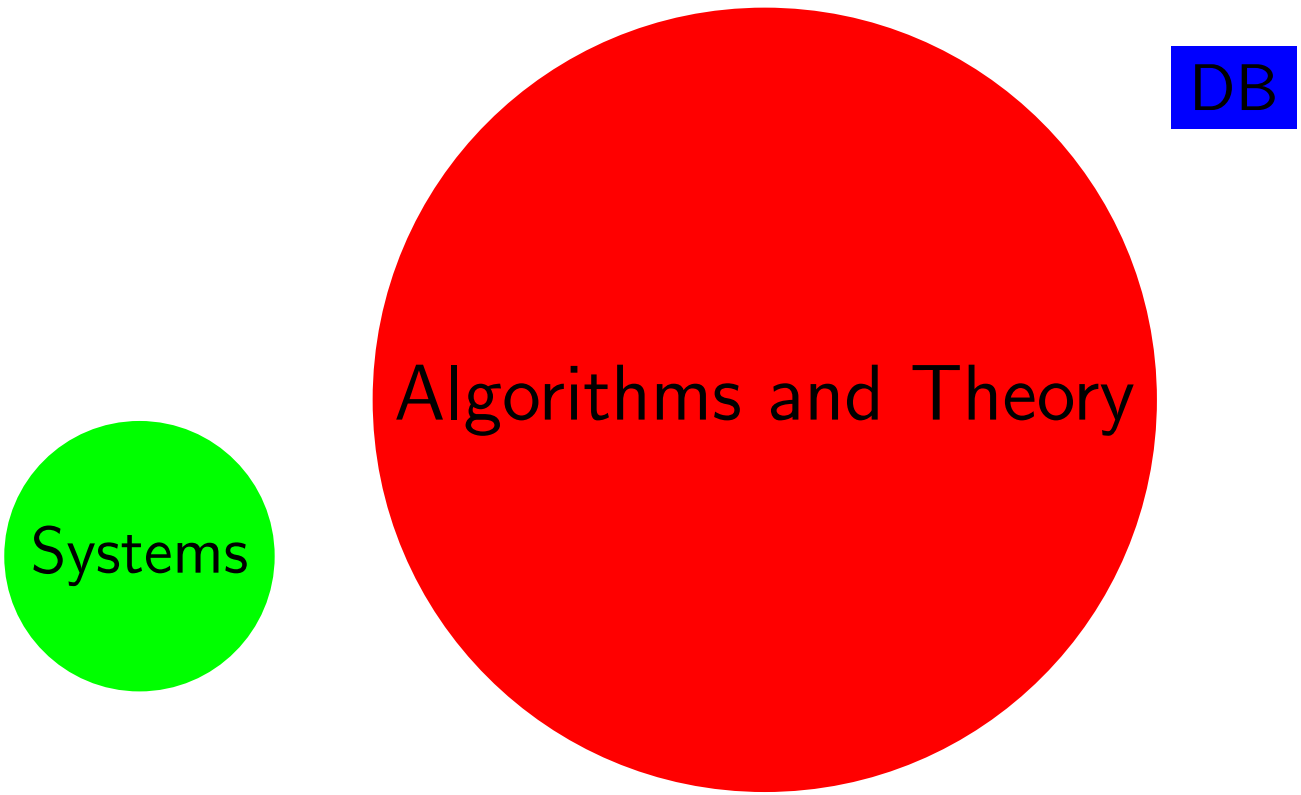
Acknowledgements: I am indebted to Michael Böhlen and Stefano Rizzi for providing me their slides, upon which these lecture notes are based.

Outline

Outline

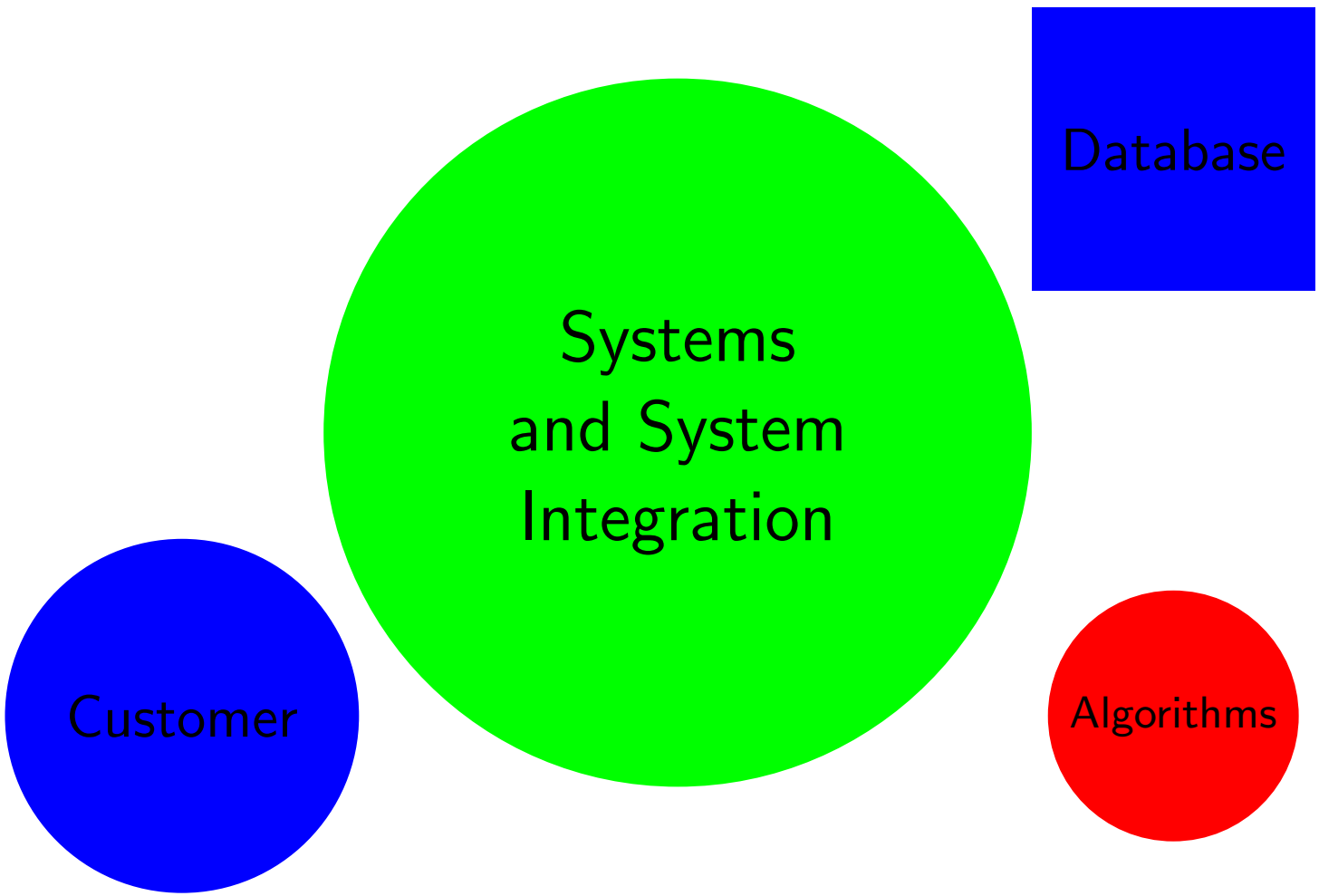
The Big Picture of Data Warehousing/1

- What is important for researchers



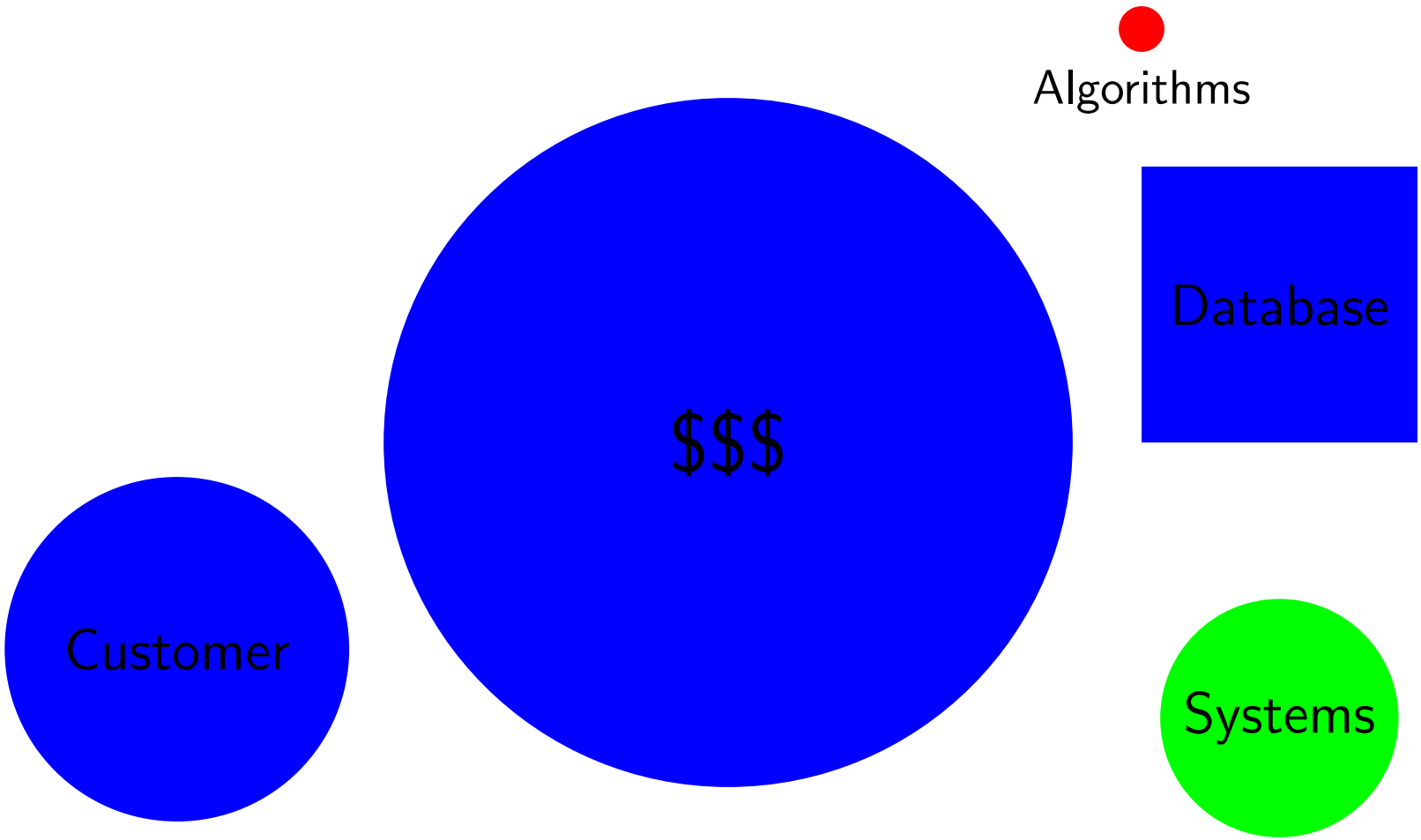
The Big Picture of Data Warehousing/2

- What is important for real world applications



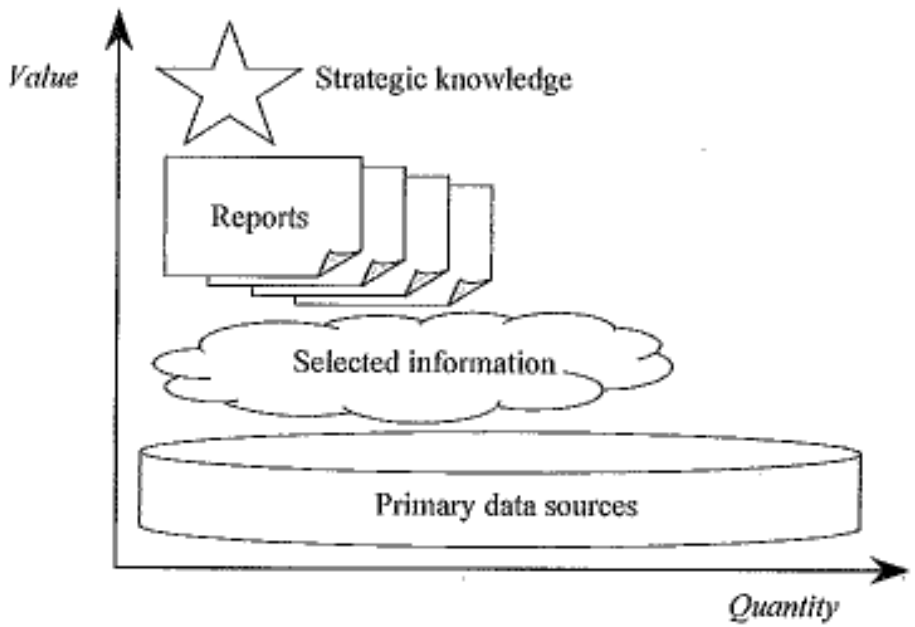
The Big Picture of Data Warehousing/3

- What is important for businesses



Information Value and Data/1

- Until the mid-1980's, enterprise databases stored only operational data
  - Data about business operations for daily management
- Today, enterprise must have quick and comprehensive access to information required for decision making.
- This strategic information is extracted from operational data stored in operational databases
  - Progressive selection and aggregation



Information Value and Data/2

- In 1996, R. Kimball summed up the needs/claims of users as follows:
  - We have heaps of data, but we cannot access it!
  - 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!

Information Value and Data/3

- 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.

What is Business Intelligence?

- **Business Intelligence (BI)** is a set of processes, tools, and technologies to transform business data into timely and accurate information to support decisional processes. BI systems include
  - Data Warehousing (DW)
  - On-Line Analytical Processing (OLAP)
  - Data Mining (DM) and Data Visualization (VIS)
  - Decision Analysis (what-if)
  - Customer Relationship Management (CRM)
- Data warehousing has also been used as a synonym for BI
- BI systems are used by decision makers to get a comprehensive knowledge of the business and to define and support their business strategies.
- The goal of BI 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.

Example BI Queries

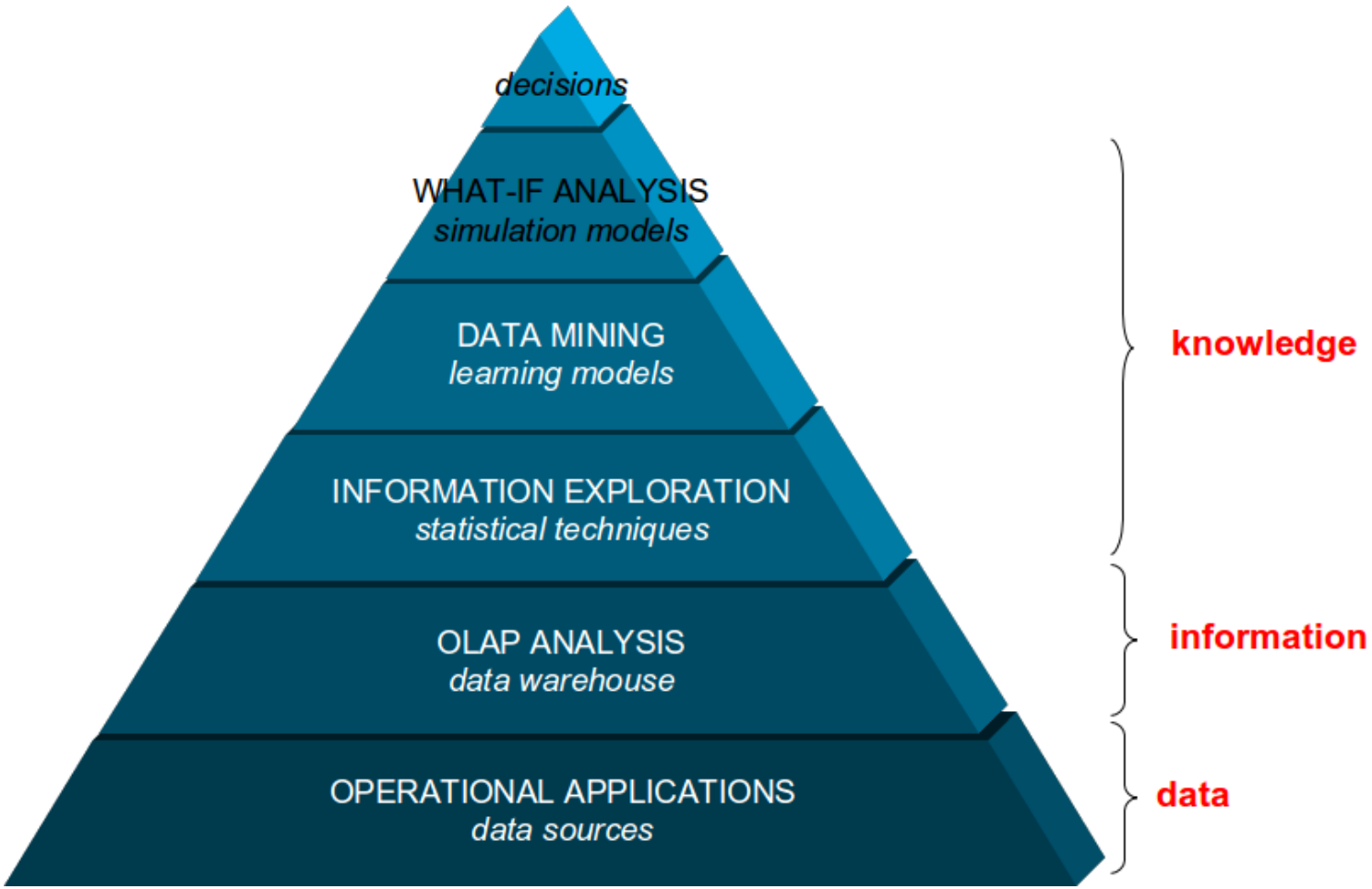
- Query Q1: On October 11, 2000, find the 5 top-selling products for each product subcategory that contributes more than 20% of the sales within its product category.

- Query Q2: As of March 15, 1995, determine shipping priority and potential gross revenue of the orders that have the 10 largest gross revenues among the orders that had not yet been shipped. Consider orders from the book market segment only.

- Regular DB models and systems are not suitable for this type of queries
  - complicated to formulate queries
  - inefficient query evaluation

  $\Rightarrow$ New models and instruments are needed!

The BI Pyramid



BI vs. Artificial Intelligence

- BI is the opposite of Artificial Intelligence (AI)
  - AI systems make decisions for the users
  - BI systems help users make the right decisions, based on the available data
  - Many BI techniques have roots in AI, though.

BI is Crucial and Growing/1

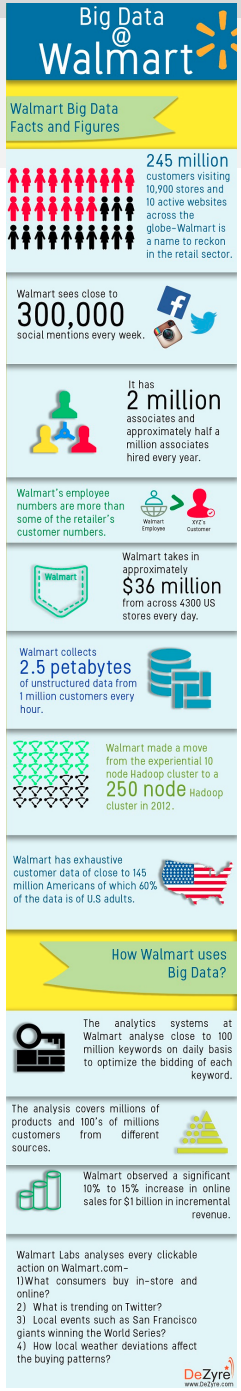
- Meta Group: DW alone = $15 Bio. in 2000
- Palo Alto Management Group: BI = $113 Bio. in 2002
- The Web made BI more necessary
  - Customers do not appear physically in the store
  - Customers can change to other stores more easily
- Thus:
  - You have to know your customers using data and BI.
  - Web logs make it possible to analyze customer behavior in more detail than before, e.g., what was not bought?
  - Combine web data with traditional customer data
- Wireless Internet adds further to this
  - Customers are always online
  - Customers position is known
  - Combine position and knowledge about customer ⇒ very valuable
    - location-based advertising

BI is Crucial and Growing/2

- Gartner, 2009:
  - Organizations will expect IT leaders in charge of BI and performance management initiatives to help transform and significantly improve their business
  - Because of lack of information, processes, and tools, through 2012, more than 35% of the top 5,000 global companies will regularly fail to make insightful decisions about significant changes in their business and markets.
  - By 2010, 20% of organizations will have an industry-specific analytic application delivered via software as a standard service of their business intelligence portfolio.
  - In 2009, collaborative decision making will emerge as a new product category that combines social software with business intelligence platform capabilities.
- S. Chaudhuri, U. Dayal, V. Narasayya, CACM 2011:
  - Today, it is difficult to find a successful enterprise that has not leveraged BI technology for their business.
- Gartner's 2012 CIO survey showed that analytics and BI is the number one technology priority for CIOs in 2012.

BI is Crucial and Growing/3

- Big Data Analysis @Walmart
  - USA's largest supermarket chain
  - Has DW with all ticket item sales
  - Uses DW and mining heavily to gain business advantages.
  - Analysis of associations within sales tickets:
    - Discovery: Beer and diapers on the same ticket.
    - Men buy diapers, and must "just have a beer".
    - Put the expensive beers next to the diapers.
    - Put beer at some distance from diapers with chips, videos in-between!
  - Wal-Mart's suppliers use the DW to optimize delivery.
    - The supplier puts the product on the shelf.
    - The supplier only get paid when the product is sold.
- Other applications: Web log mining
  - What is the association between time of day and requests?
  - What user groups use my site?
  - How many requests does my site get in a month? (Yahoo)



Remarks about the Data Warehouse Part

- We learn how to design, build, and use a data warehouse.
- Relevance to the real world is an important guideline.
- Not only/mainly crisp algorithms, theorems, etc.
- We will look at a number of concrete and important case studies.
- A good way to prepare and learn the subject is to participate to lectures.
- Data mining is taught in a different course.

Content of the Data Warehouse Part

- Data warehousing: business intelligence, data integration, data warehouse, facts, dimensions, DW design
- ETL and advanced modeling: ETL process, handling changes in dimensions
- SQL OLAP extensions: analytical functions, crosstab, group by extensions, hierarchical cube, moving windows
- Generalized multi-dimensional join: GMDJ, evaluation, subqueries, optimization rules, distributed evaluation
- DW performance: pre-aggregation, lattice framework, view selection, view maintenance, bitmap indexing

Literature

- Matteo Golfarelli, Stefano Rizzi. *Data Warehouse Design: Modern Principles and Methodologies*. McGraw-Hill, 2009. (recommended!)
- Alejandro A. Vaisman, Esteban Zimnyi. *Data Warehouse Systems: Design and Implementation*. Series: Data-Centric Systems and Applications Springer, 2014.
- Ralph Kimball, Margy Ross. *The Data Warehouse Toolkit*, 2nd edition.
- William H. Inmon, *Building the Data Warehouse*, 4th edition.
- Selected research papers will be announced later.

Outline

BI: Key Problems

1. Complex and unusable models in operational systems
   - Many DB models are difficult to understand
   - DB models do not focus on a single clear business purpose
2. Same data found in many different systems
   - Examples: customer data in many different systems, residential address of citizens in many public administration DBs, etc.
   - The same concept is defined and stored differently
3. Data is suited only for operational systems
   - Accounting, billing, etc.
   - Do not support analysis across business functions
4. Data quality is bad
   - Missing data, imprecise data, different use of systems
5. Data are volatile
   - Data deleted in operational systems (6 months)
   - Data change over time no historical information

BI: Solution

- A new analysis environment with a **data warehouse** at the core, where data is
  - integrated (logically and physically),
  - subject oriented (versus function oriented),
  - supporting management decisions (different from organization),
  - stable (data is not deleted, several versions),
  - time variant (data can always be related to time).

Definition of a Data Warehouse [Barry Devlin, IBM]

- A data warehouse is simply a
  - single,
  - complete, and
  - consistent

  store of data obtained from a variety of sources and made available to end users in a way they can understand and use it in a business context.

Definition of a Data Warehouse [William H. Inmon]

- A data warehouse is a
  - subject-oriented,
  - integrated,
  - time-varying, and
  - non-volatile

  collection of data that is used primarily in organizational decision making.

Requirements for a DW Architecture

- Separation: Analytical and transaction processing should be kept apart as much as possible.
- Scalability: HW and SW should be easy to upgrade as the data volume and the number of user requirements increase.
- Extensibility: Should be possible to host new applications and technologies without redesigning the whole system.
- Security: Monitoring accesses is essential because of the strategic data stored in DW.
- Administerability: Administration not too difficult.

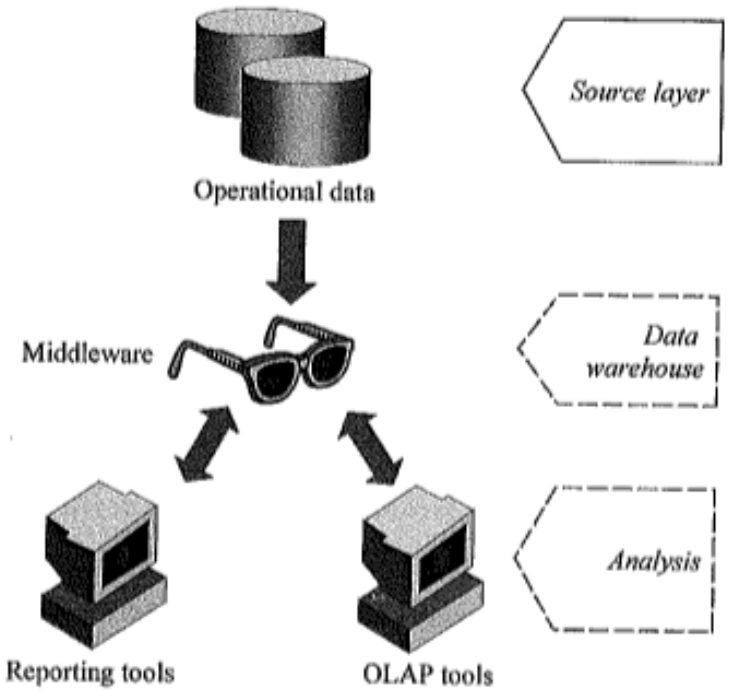
Single-layer DW Architecture [Golfarelli & Rizzi]

- Only source layer is physical
- DW exists only virtually as view
- Not frequently used in practice

\+ Mimimizes amount of stored data

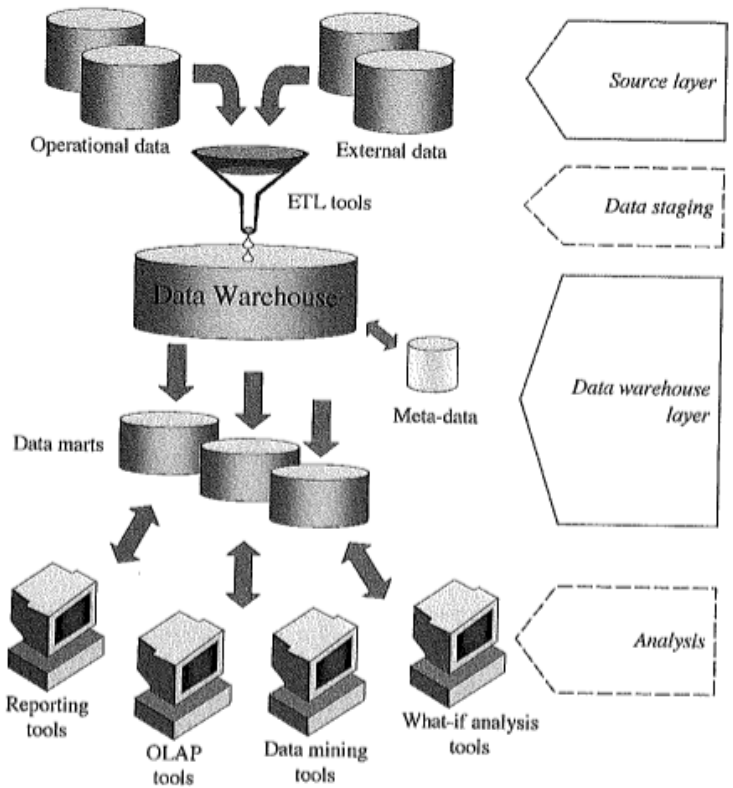
– No separation between analytical and transactional processing, hence queries affect regular workload

– No additional data can be stored



Two-layer DW Architecture [Golfarelli & Rizzi]

- Source layer and DW exist physically
  $\Rightarrow$ clear separation
- Data staging: extraction, transformation, and cleaning of data
- (Primary) DW can be source for data marts



\+ Clearly separates analytical and transactional processing, hence queries do not affect regular workload

\+ DW is structured according to multidimensional model

\+ DW is accessible, even if source systems are unavailable

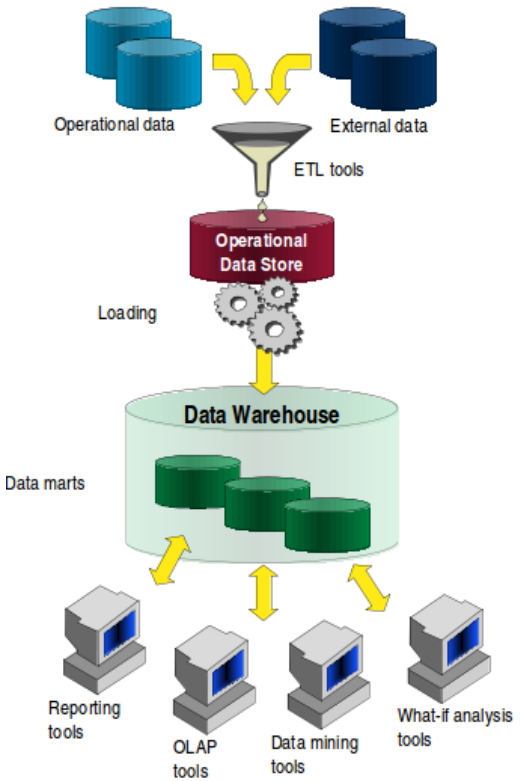
Data Mart (DM)

- A data mart is a subset or an aggregation of the data stored in a primary DW.
- It includes a set of information pieces relevant to a specific business area, corporate department, or category of users.
- DMs are typically populated from a primary DW (dependent DM)
- Might also be populated direclty by data sources (independent DM)
- Used as building blocks in incremental DW design
- Can deliver better performance

Three-layer DW Architecture [Golfarelli & Rizzi]

- Reconciled layer (in addition to source and DW layer): Materialization of integrated, clean and consistent operational data

+ Reconciled layer is common reference data model for whole enterprise, i.e., single, detailed, comprehensive, and top-quality data source
+ Separates source data extraction from DW population
– More data redundancy



EXTRACTION, TRANSFORMATION, AND LOADING:

ETL processes extract data from sources, transform and clean them, and finally load them in the ODS and in the data warehouse

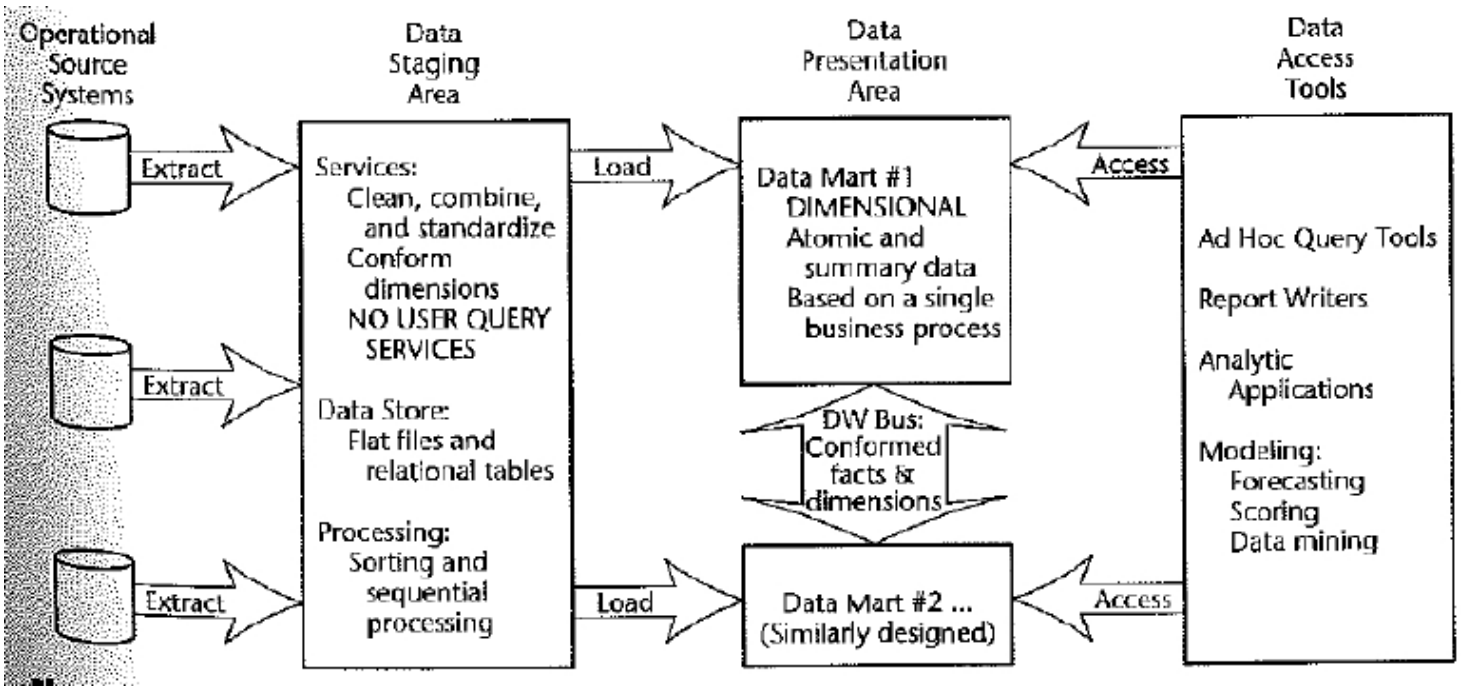
OPERATIONAL DATA STORE:

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

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

Three-layer DW Architecture [Kimball]



Data Integration

- Two different ways to integrate the data from the sources:
  - Query-driven
  - Warehouse-driven

Query-driven Data Integration

- Data is integrated on demand (lazy)
- Corresponds to single-layer architecture
- PROS
  - Access to most up-to-date data (all source data directly available)
  - No duplication of data
- CONS
  - Delay in query processing due to
    - slow (or currently unavailable) information sources
    - complex filtering and integration
  - Inefficient and expensive for frequent queries
  - Competes with local processing at sources
  - Data loss at the sources (e.g., historical data) cannot be recovered
- Has not caught on in industry

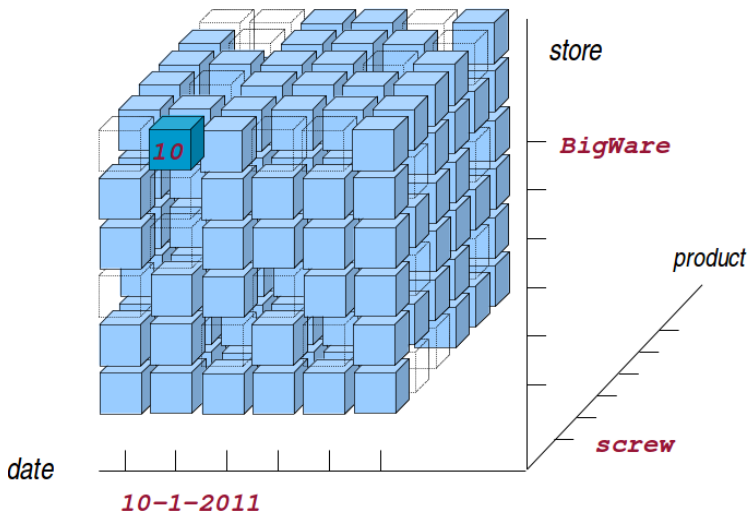
Warehouse-driven Data Integration

- Data is integrated in advance (eager)
- Data is stored in DW for querying and analysis
- PROS
  - High query performance
  - Does not interfere with local processing at sources
    - Assumes that DW update is possible during downtime of local processing
    - Complex queries are run at the DW
    - OLTP queries are run at the source systems
- CONS
  - Duplication of data
  - The most current source data is not available
- Has caught on in industry

Outline

Multidimensional Model/1

- Key for representing and querying information in a DW.
- Stores information about enterprise-specific facts that affect decisions
  - The occurrence of a fact is often called event
- Facts are characterized by
  - measures: numerical values that provide a quantitative description of events
  - dimensional attributes: provide different perspectives for analyzing the facts
- (Data) Cube: metaphor to store facts/events in an n-dimensional space
  - cells store measures
  - axes are the dimensions
- Example: data cube for sales facts



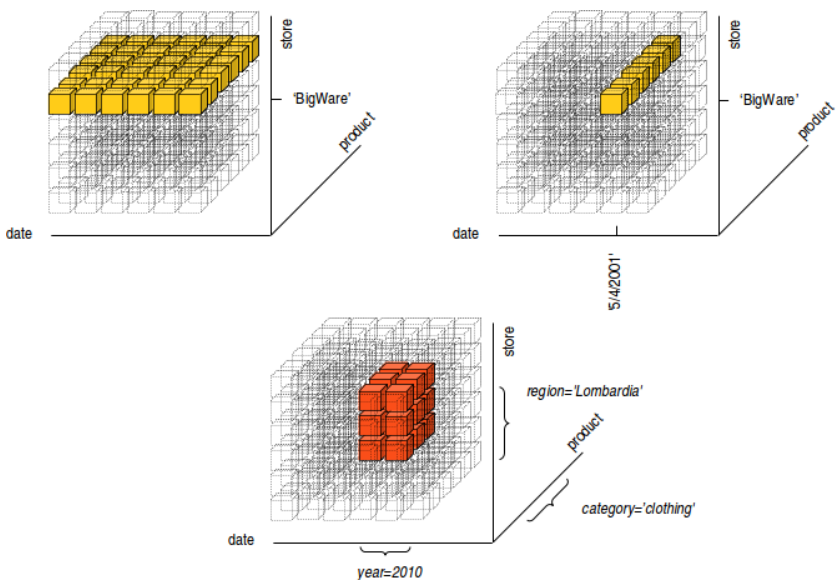
Multidimensional Model/2

- Multidimensional model is very easy to use and understand (main reason for success)
- Examples of typical queries in DW
  - What is the total amount of receipts recorded last year per state and per product category?
  - What is the relationship between the trend of PC manufacturers' shares and quarter gains over the last five years?
  - Which orders maximize receipts?
  - What is the relationship between profit gained by the shipments consisting of less than 10 items and the profit gained by the shipments of more than 10 items?
- There are a number of operators that allow to answer such queries easily ($\rightarrow$ OLAP)

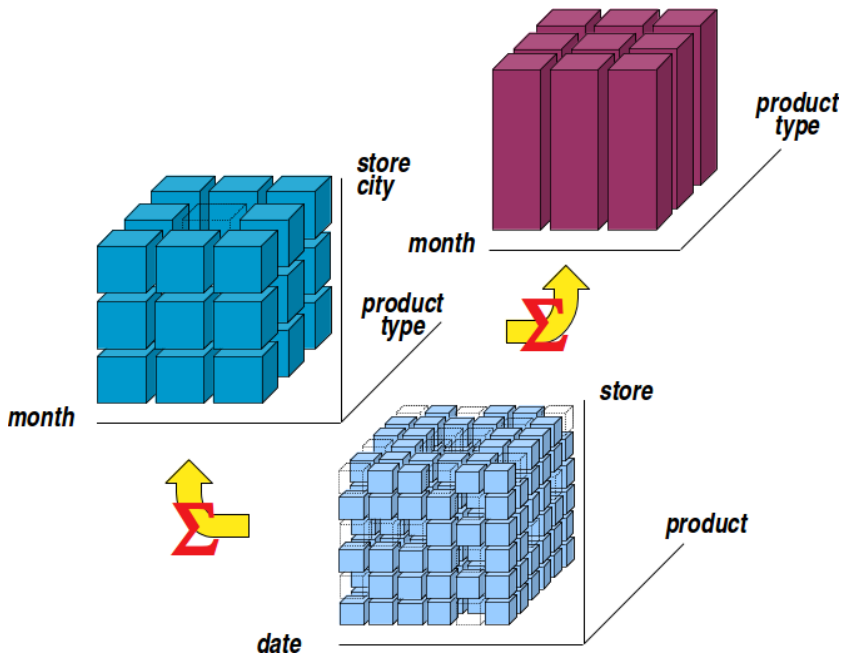
Online Analytical Processing (OLAP)

- OLAP is an approach to answer multi-dimensional analytical queries swiftly
  - Essentially aggregations along different dimensions

- Slicing and dicing



- Aggregation



OLTP versus OLAP/1

- Different query types in operational systems and DW
  - OLTP in operational DB
  - OLAP in DW
- On-Line Transaction Processing (OLTP)
  - Many "small" queries on a small number of tuples from many tables that need to be joined
  - Frequent updates
  - The system is always available for both updates and reads
  - Smaller data volume (few historical data)
  - Complex data model (normalized)
- On-Line Analytical Processing (OLAP)
  - Fewer, but "bigger" queries that typically need to scan a huge amount of records and doing some aggregation
  - Frequent reads, in-frequent updates (daily, weekly)
  - 2-phase operation: either reading or updating
  - Larger data volumes (collection of historical data)
  - Simple data model (multidimensional/de-normalized)

OLTP versus OLAP/2

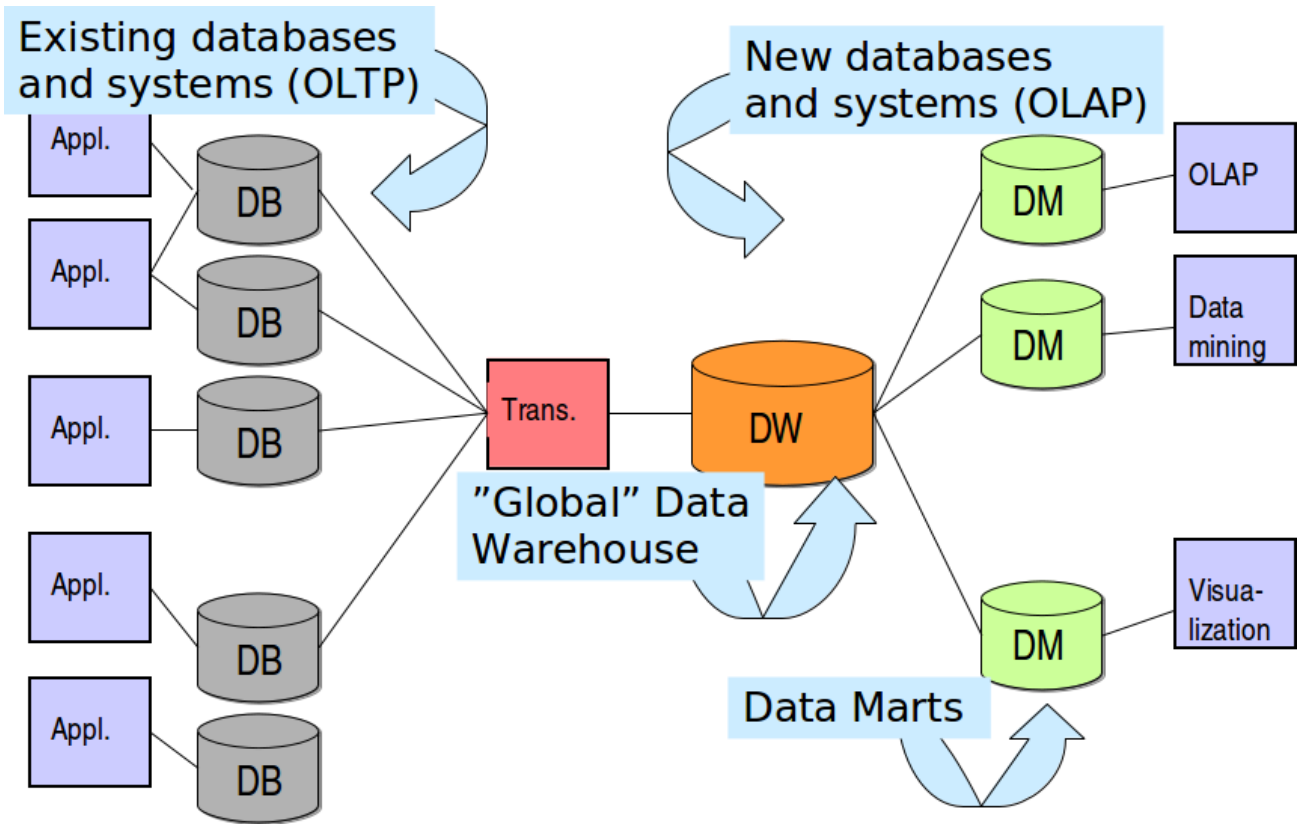
- A mix of analytical queries (OLAP) with transactional routine queries (OLTP) inevitably slows down the system
- This does not meet the needs of users of both types of queries
- Separate OLAP from OLTP by
  - Creating a new repository (DW) that integrates data from various sources
  - Makes data available for analysis and evaluation aimed at decision-making processes

OncoNet Example

- OncoNet is a (small) system for the management of patients undergoing a cancer therapy used in the Hospital of Meran
  - $>$ 200 tables
- Well-suited for daily management of patients
- But: statistical analysis are expensive
  - takes up to 12 hours
  - tables are locked for that time
  - run queries over weekend
- A DW approach reduced the runtime of the same queries to a few seconds (BSc-thesis of A. Heinisch)

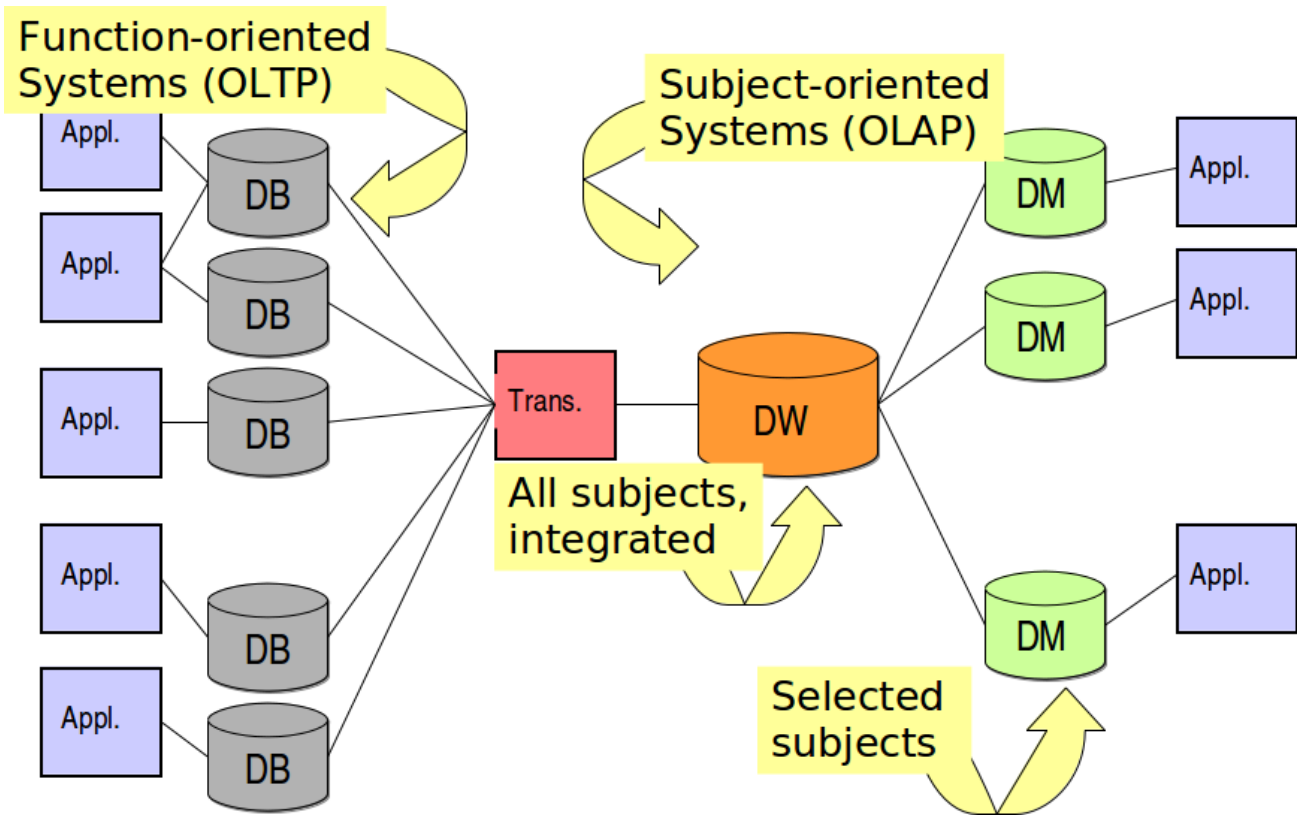
		Testo	Numero
☑	Q	STAO GENERALE (ECOG)	650
☑	O	Attivita' normale	369
☑	O	Attivita' ridotta, non allettato, lavora	203
☑	O	Non e' in grado di lavorare; richiede assistenza; < 50% allettato	58
☑	O	Unfähig sich selbst zu versorgen; kontinuirliche Pflege oder Hosp	17
☑	O	100 % allettato	1
☑	Q	FATIGUE	650
☑	O	No	366
☑	O	Letargia	105
☑	O	Fatigue moderata	110
☑	O	Fatigue gravert	66
☑	O	Allettato	1
☐	Q	INSONNIA	
☐	X	Problemi di addormentarsi	

OLTP versus OLAP/3



OLTP versus OLAP/4

- OLTP is function-oriented, OLAP subject-oriented



Summary BI

- Business Intelligence (BI) is well-recognized and is a combination of a number of tools and techniques to support critical decision making in businesses.
- BI systems provide a comprehensive knowledge of the business and enable data-/evidence-based decisions
- BI helps to transform data into strategic knowledge
- Data Warehouse (DW) is at the core of BI
- BI is crucial for any business, and it is growing

Summary DW

- A Data Warehouse (DW)
  - is at the core of BI;
  - provides a complete, consistent, subject-oriented and time-varying collection of the data;
  - provides comprehensive knowledge about your business;
  - allows to separte OLAP from OLTP.
- A good DW is a prerequisite for BI, but a DW is a means rather than a goal . . . it is only a success if it is heavily used.
- Single-, two-, and three-layer architectures of DWs
- Separate OLAP from OLTP by creating a DW