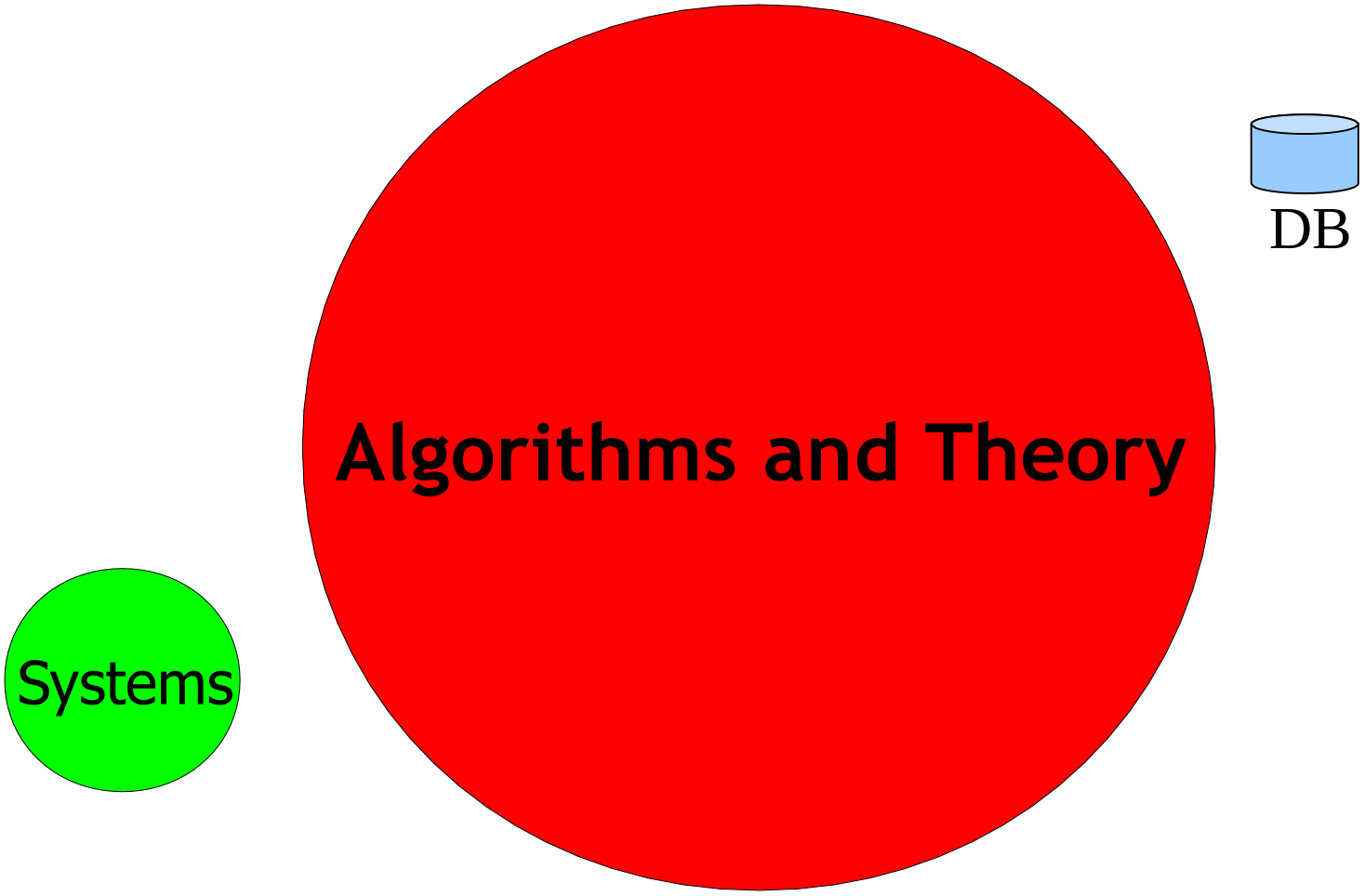
Data Warehousing and Data Mining
– Introduction –

- ✓ General introduction to DWDM
- ✓ Business intelligence
- ✓ OLTP vs. OLAP Data
- ✓ integration Methodological
- ✓ framework
- ✓ DW definition

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

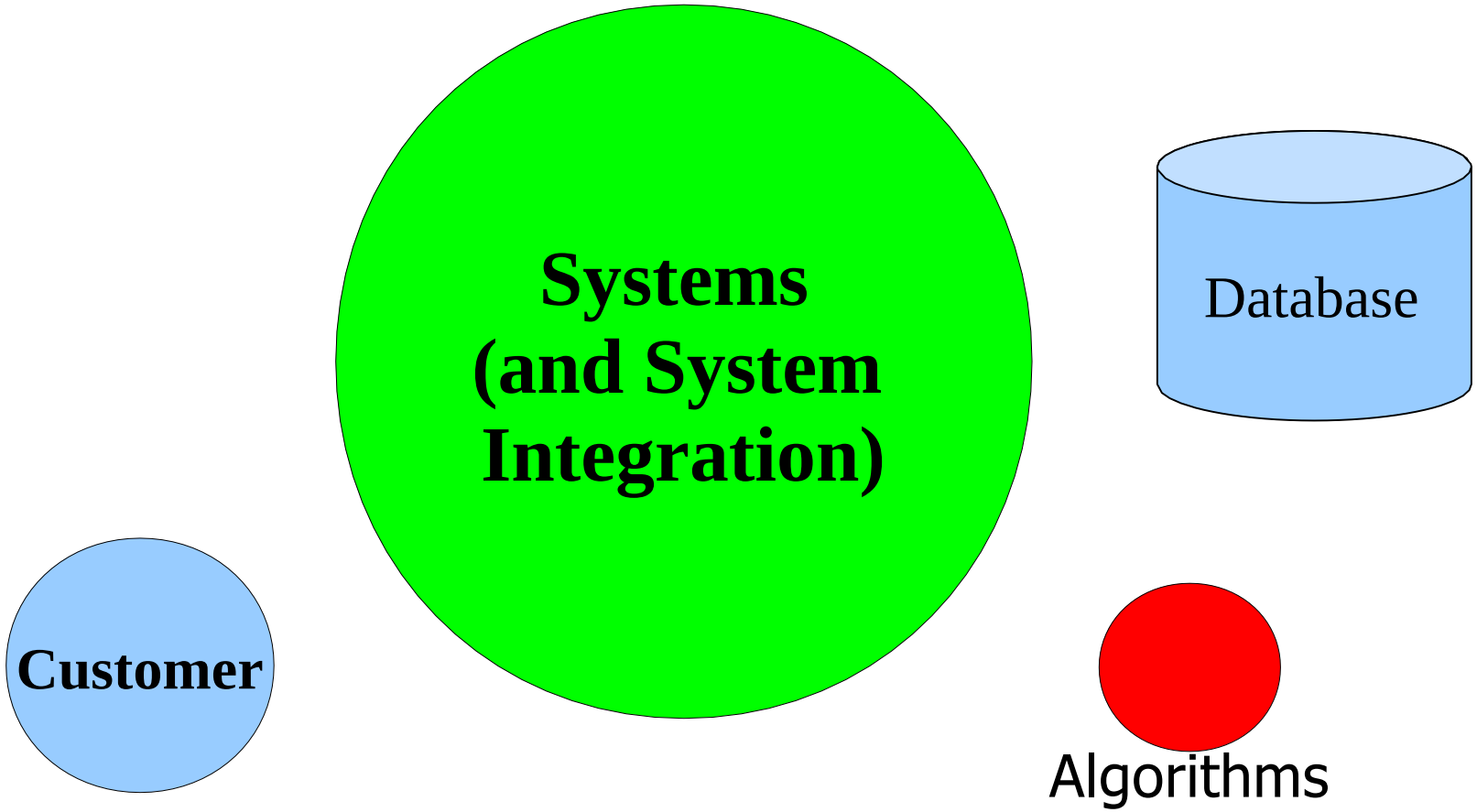
The Big Picture of DWDM/I

- What's important for researchers:



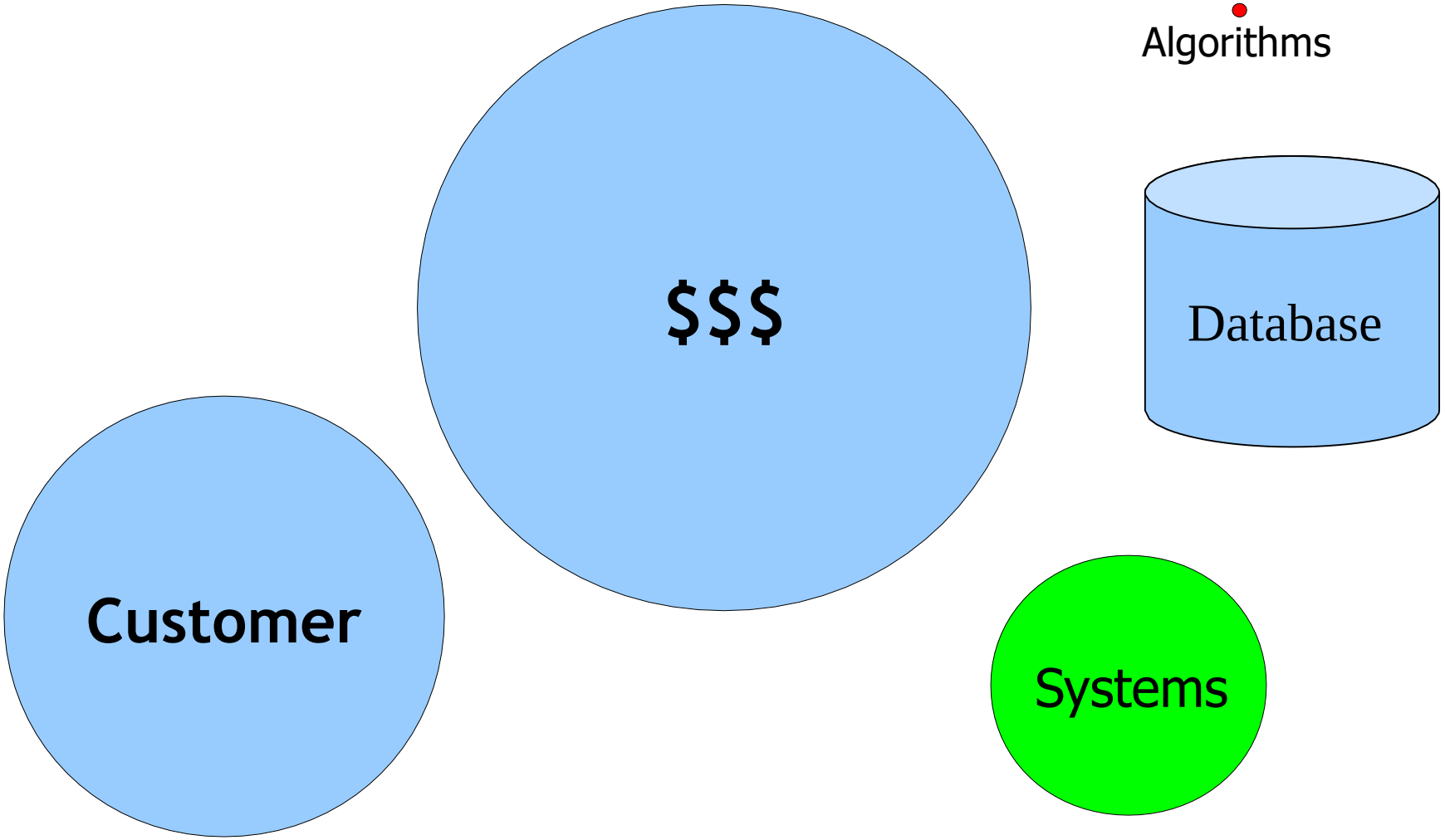
The Big Picture of DWDM/2

- What's important for real world applications:



The Big Picture of DWDM/3

- What's important for businesses:



Computer Science and Decision Making

- 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.**

Remarks about the DW 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.

Content of the DW Part

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

What is Business Intelligence?/1

- BI is a set of processes, tools, and technologies to transform business data into timely and accurate **information to support decisional processes**
  - **Data Warehousing** (DW)
  - On-Line Analytical Processing (OLAP)
  - Data Mining (DM) and Data Visualization (VIS)
  - Decision Analysis (what-if)
  - Customer Relationship Management (CRM)
- 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 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**.

What is Business Intelligence?/2

- 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.

The BI Pyramid

Example BI Queries

- **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.
- **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 database models and systems are not suitable for this type of queries.

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 is possible to analyze customer behavior in more detailed than before (what was **not** bought?)
  - Combine web data with traditional customer data
- Wireless Internet adds further to this:
  - Customers are always ”online”
  - Customer’s position is known
  - Combine position and knowledge about customer => very valuable

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.

BI: Key Problems

1) **Complex and unusable models**
   - 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**
   - Example: customer data in many different systems
   - The same concept is defined differently
3) **Data is suited 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 organization)
  - Stable (data is not deleted, several versions)
  - Time variant (data can always be related to time)

Definition of a Data Warehouse/ I

- Barry Devlin, IBM Consultant

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/2

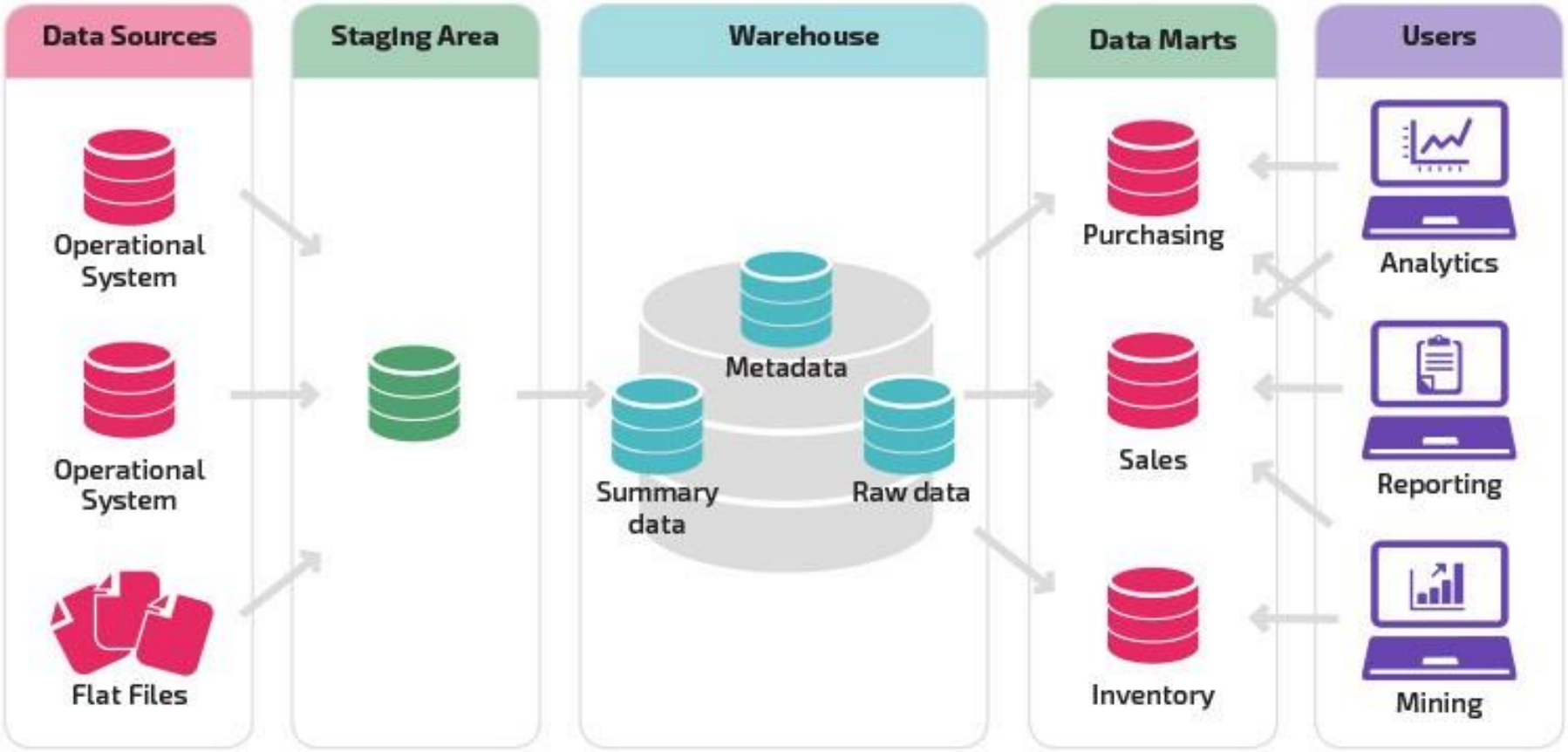
- W. H. Inmon, Building the Data Warehouse

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

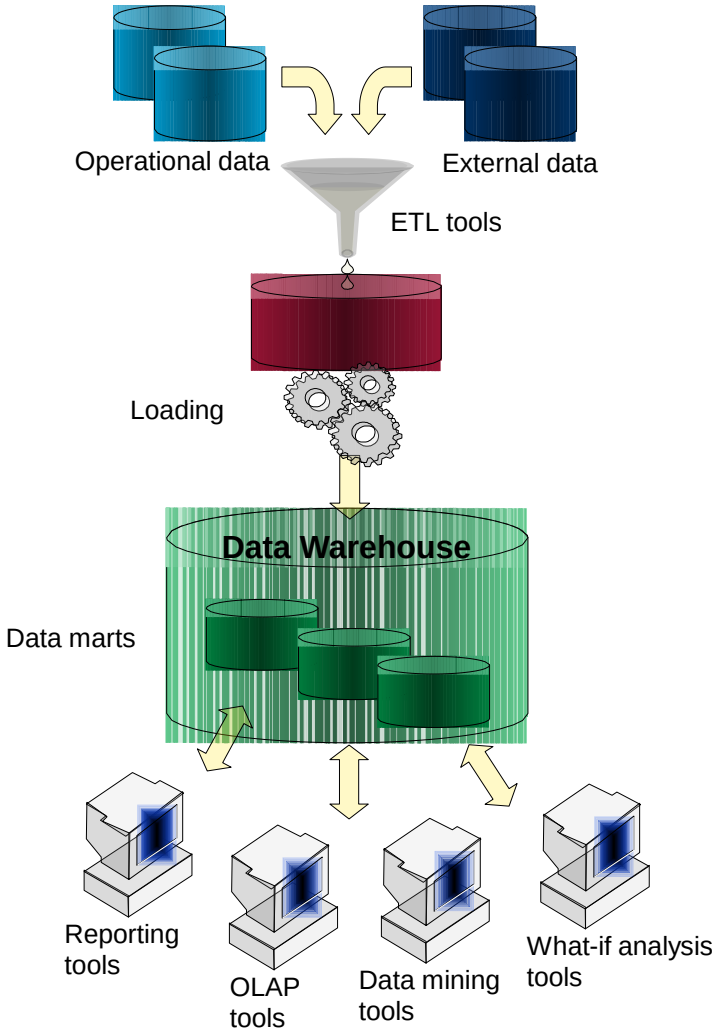
collection of data that is used primarily in
- **organizational decision making**.

DW Architecture/ I

- Basic elements of a Data Warehouse environment



DW Architecture/2



Source layer

Data staging

Reconciled layer

Data warehouse layer

Analysis

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

Data Integration

Problem:

- Different interfaces
- Different data representations
- Duplicated information
- Inconsistent information

Integrated System

Goal:

- Collect and combine information
- Provide an integrated view
- Provide a uniform user interface
- Support sharing of data

Query-Driven Data Integration

- Data is integrated **on demand** (lazy)
- PROS
  - Access to most up-to-date data (all source data directly available)
  - No duplication of data
- CONS
  - Delay in query processing
    - 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 data warehouse update is possible during downtime of local processing
    - Complex queries are run at the data warehouse
    - 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

OLTP versus OLAP/1

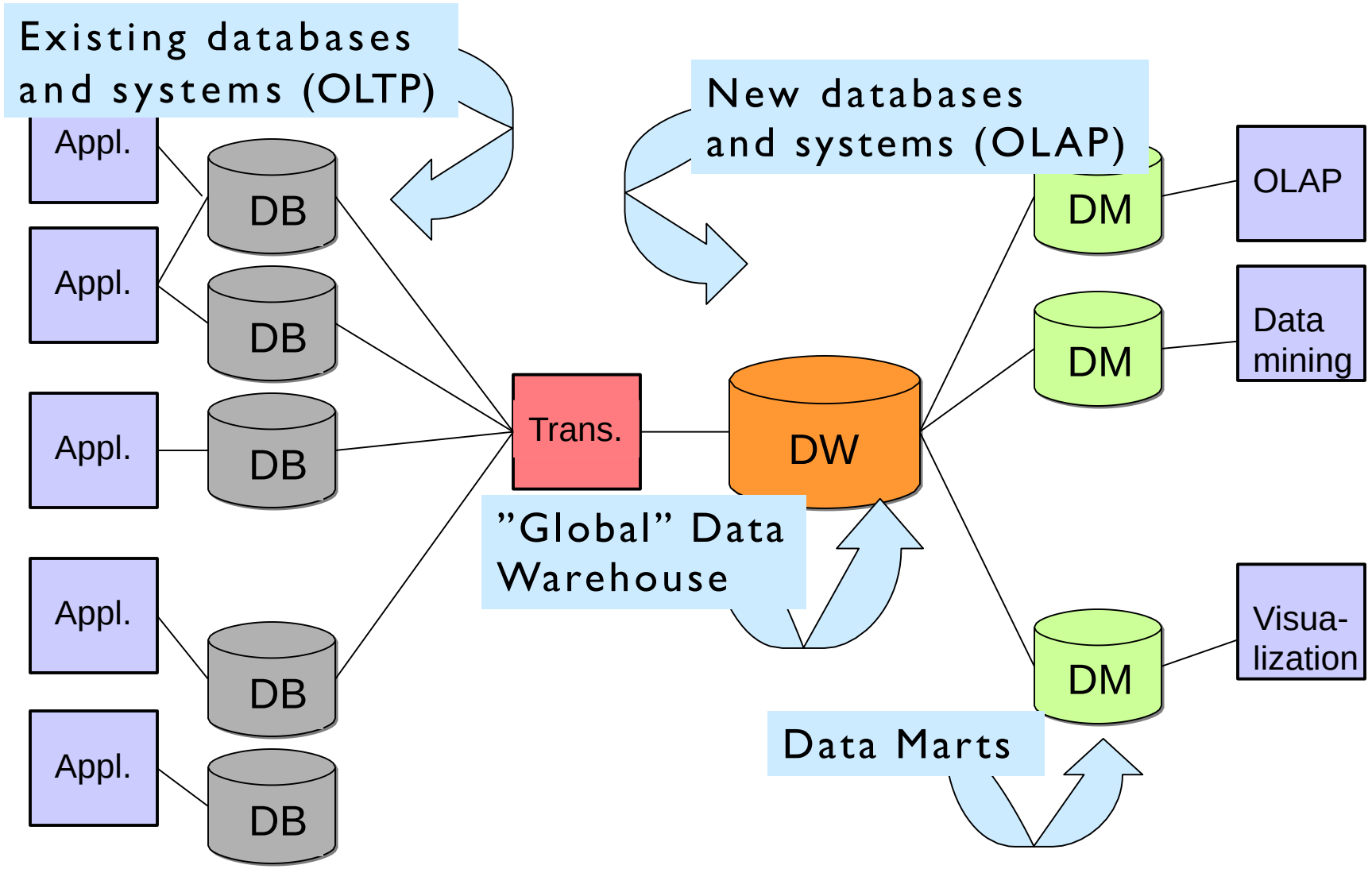
- On-Line Transaction Processing (OLTP)
  - Many ”small” queries on a small number of tuples from many tables that need to be joined
  - Frequent updates
  - The system is always available for both updates and reads
  - Smaller data volume (few historical data)
  - Complex data model (normalized)

- 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)

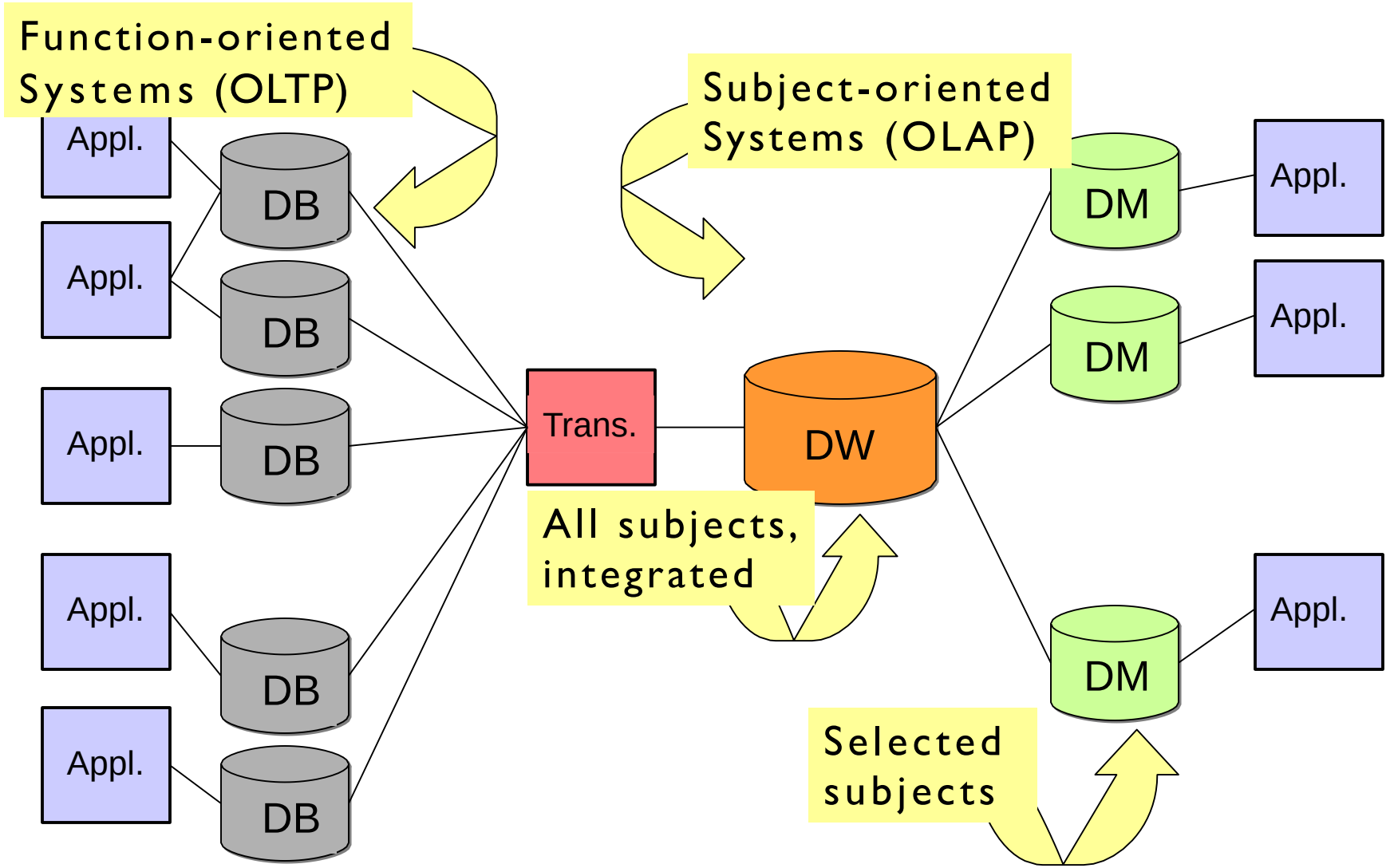
OLTP versus OLAP/2

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

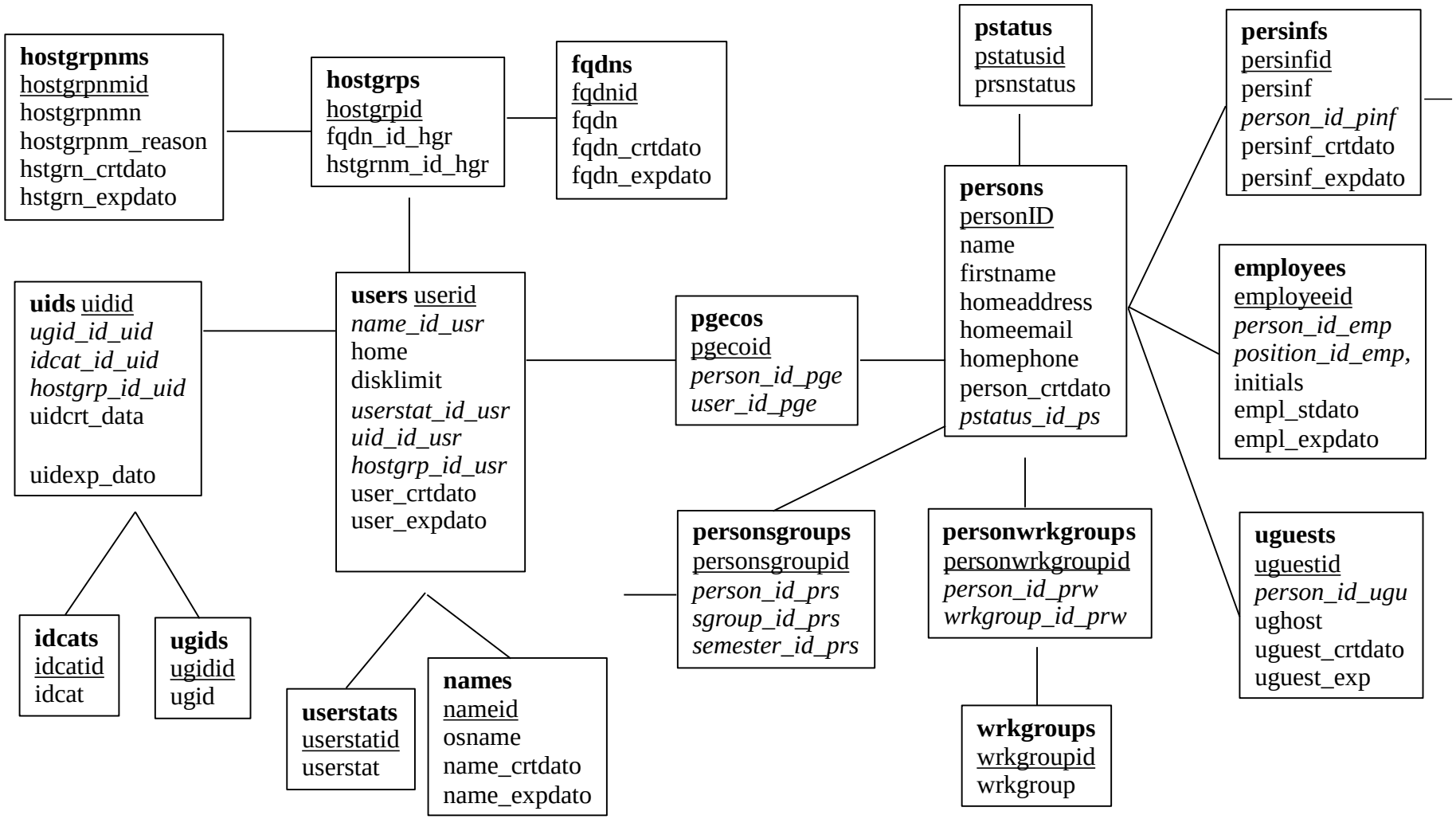
OLTP versus OLAP/3



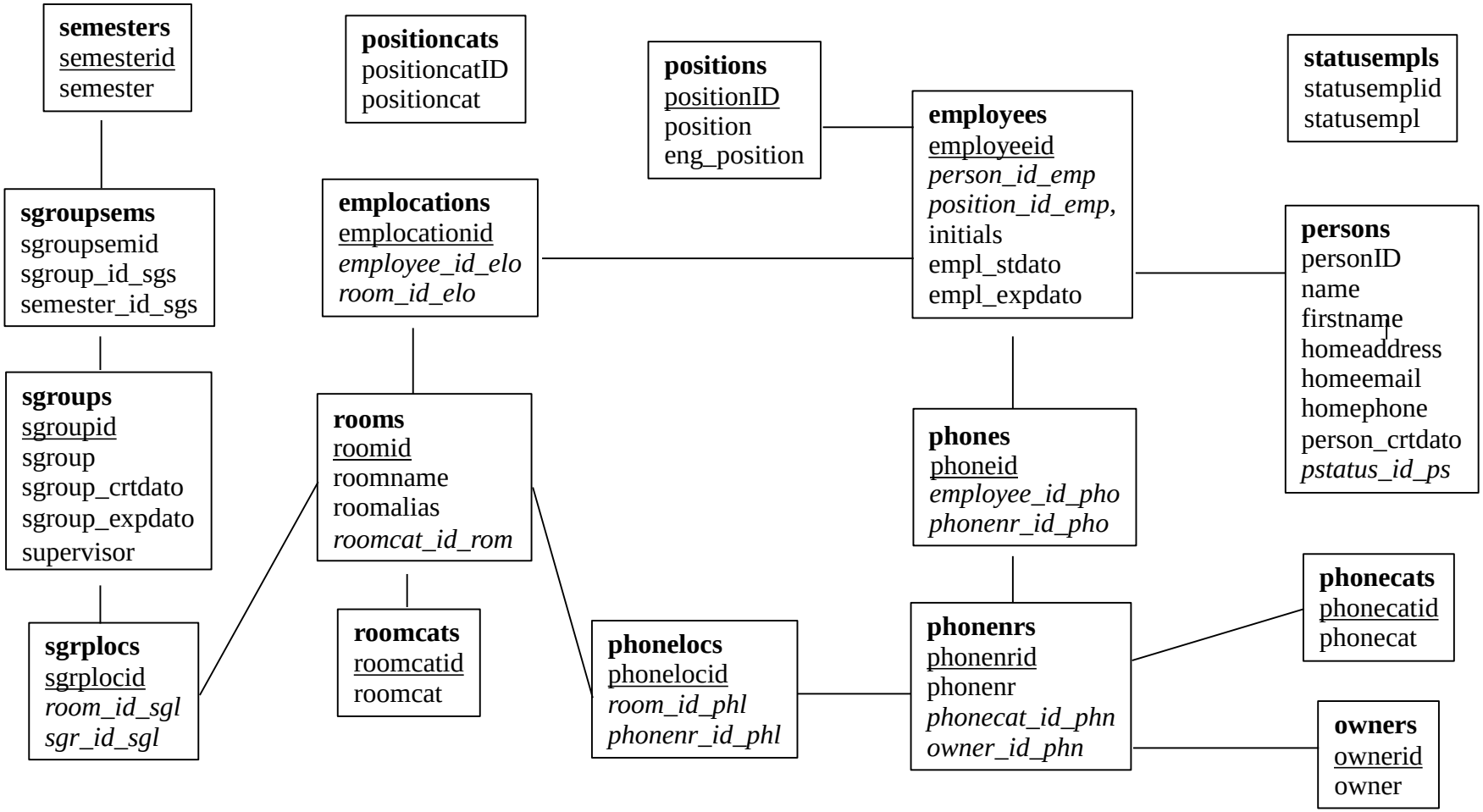
OLTP versus OLAP/4



OLTP Example: CS Dept/1

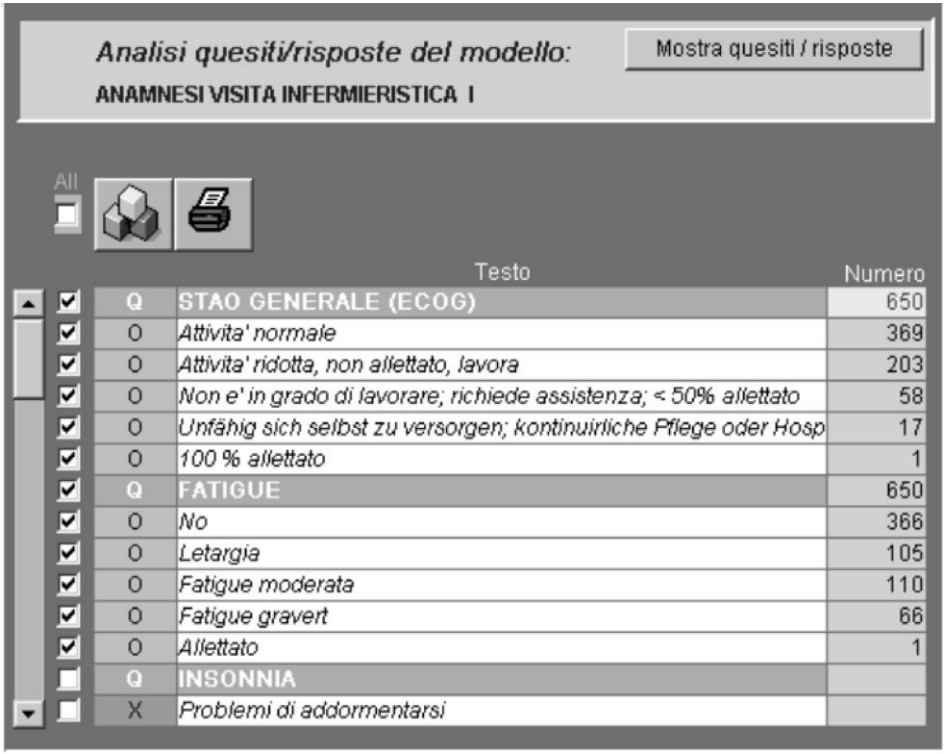


OLTP Example: CS Dept/2



OLTP Example: OncoNet

- OncoNet is a system for the management of patients undergoing a cancer therapy
  - > 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

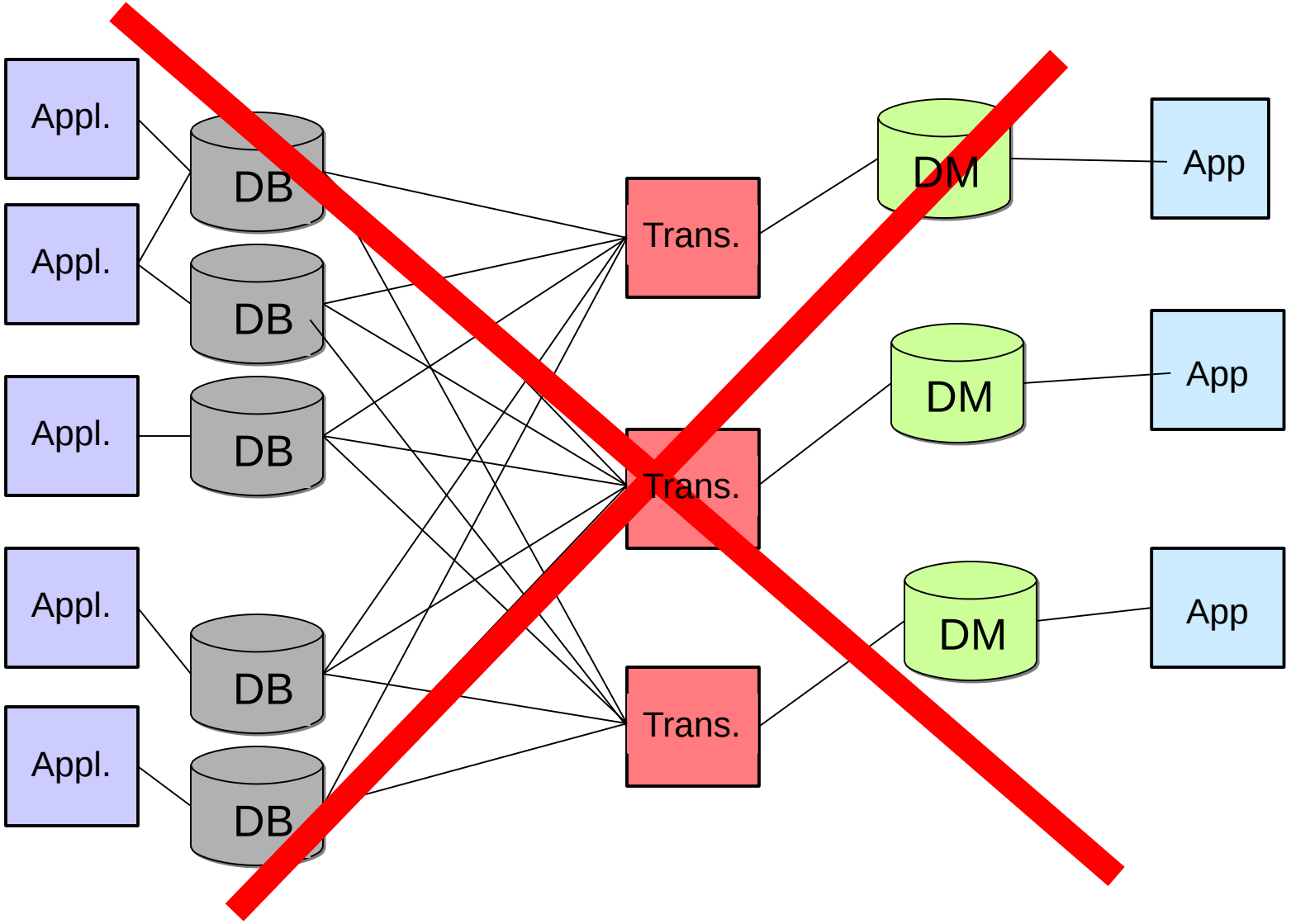


- A DW approach reduced the runtime of the same queries to a few seconds (BSc-thesis of A. Heinisch)

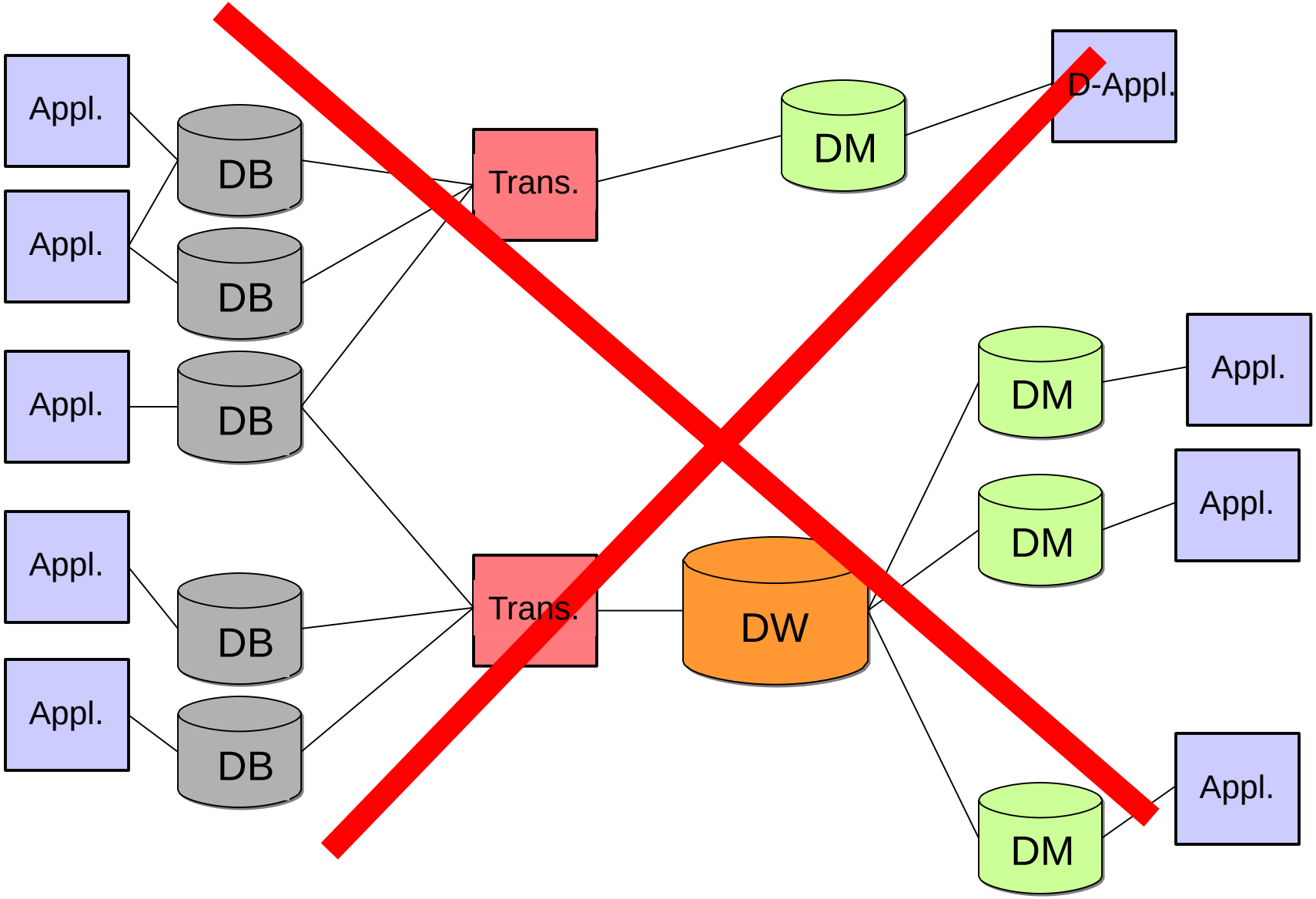
Methodological/Design Framework

- Building a DW is a **very complex task**, which requires an **accurate planning** aimed at devising satisfactory answers to organizational and architectural questions.
- A large number of organizations **lack experience and skills** that are required to meet the challenges involved in DW projects.
- The reports of DW project failures state that a major cause lies in the **absence of a global view** of the design process: in other terms, in the **absence of a design methodology.**
- Methodologies are created by closely studying similar experiences and **minimizing the risks for failure** by basing new approaches on a constructive analysis of the mistakes made previously.

Many Ways not to Do/1



Many Ways not to Do/2



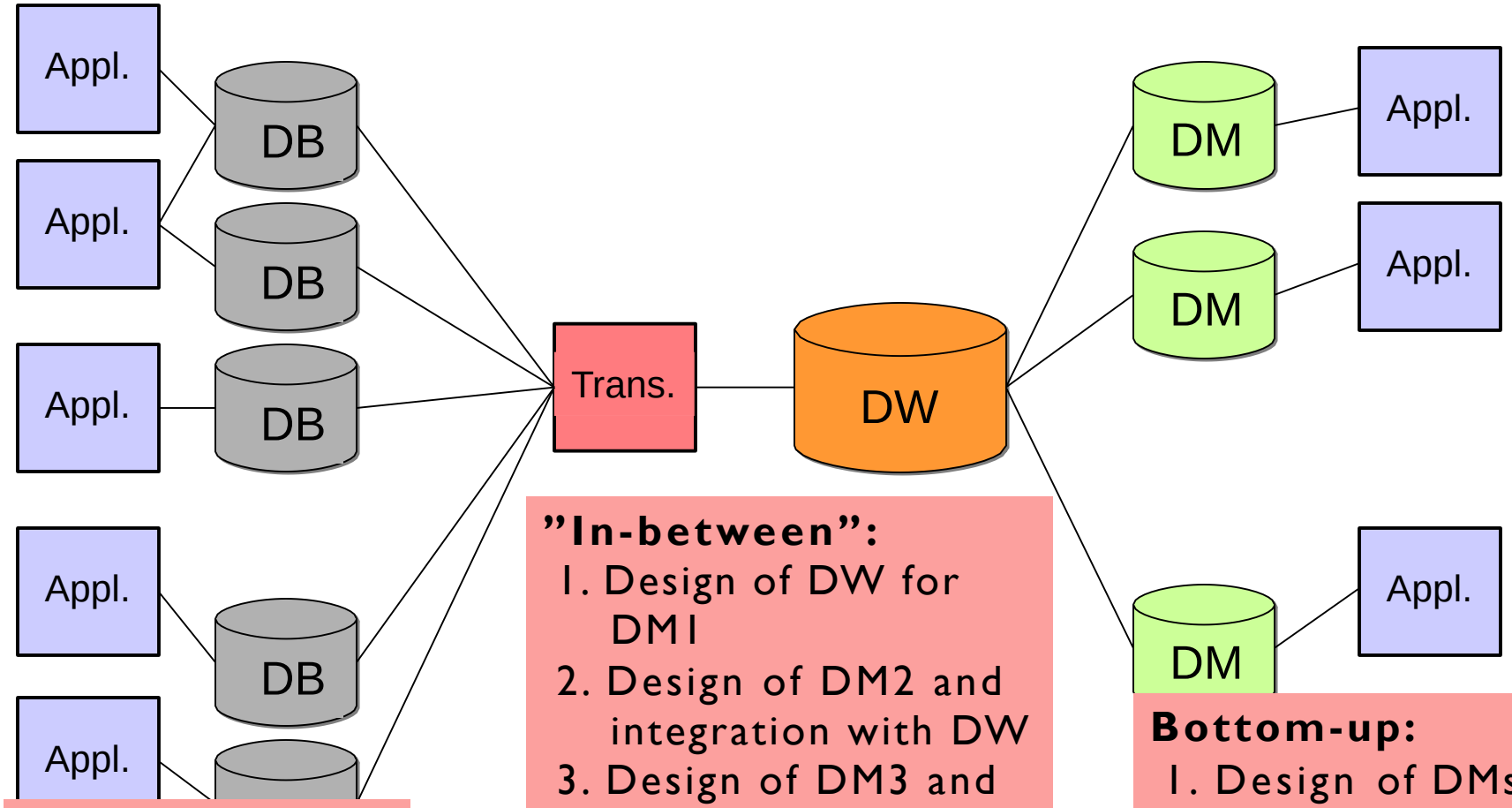
Top-down Approach

- Analyze global business needs, plan how to develop a data warehouse, design it, and implement it as a whole
  - This procedure is promising: it is **based on a global picture** of the goal to achieve, and in principle it ensures consistent, well integrated data warehouses.
  - High-cost estimates with **long-term implementations discourage** company managers from embarking on these kind of projects.
  - Analyzing and integrating all relevant sources at the same time is a **very difficult task**, even because it is not very likely that they are all available and stable at the same time.
  - It is **extremely difficult to forecast** the specific needs of every department involved in a project, which can result in the analysis process coming to a standstill.
  - Since **no working system is going to be delivered in the short term,** users cannot check for this project to be useful, so they lose trust and interest in it.

Bottom-up Approach

- DWs are incrementally built and several data marts are iteratively created. Each data mart is based on a set of facts that are linked to a specific department and that can be interesting for a user group
  - Leads to concrete **results in a short time**
  - Does **not require huge investments**
  - Enables designers to investigate **one area at a time**
  - Gives managers a **quick feedback** about the actual benefits of the system being built
  - Keeps the interest for the project constantly high
  - May determine a partial vision of the business domain

Top-down vs. Bottom-up Approach

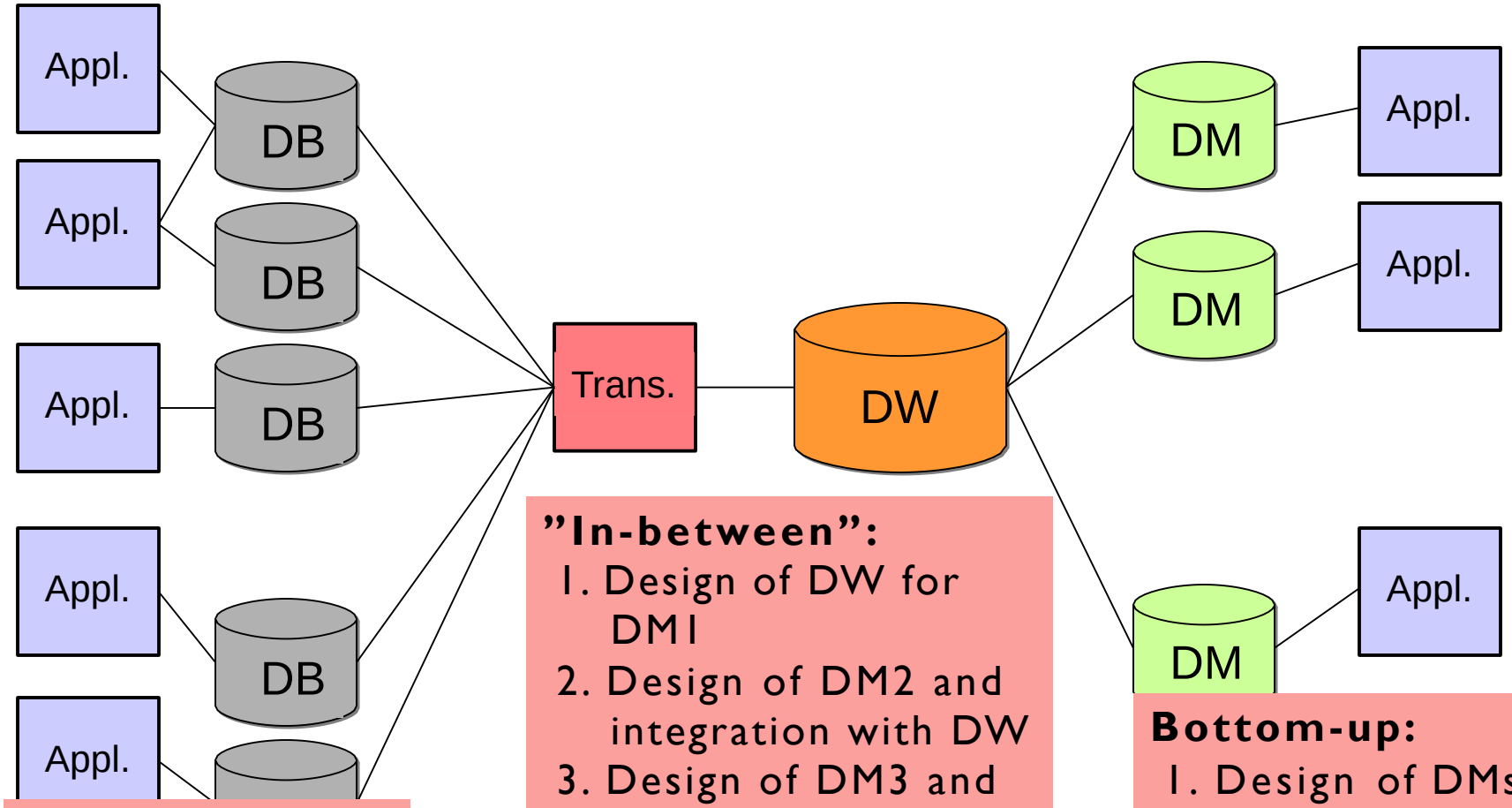


Top-down:
1. Design of DW
2. Design of DMs

"In-between":
1. Design of DW for DM1
2. Design of DM2 and integration with DW
3. Design of DM3 and integration with DW
4. ...

Bottom-up:
1. Design of DMs
2. Integration of DMs in DW
3. Maybe no physical DW

Top-down vs. Bottom-up Approach

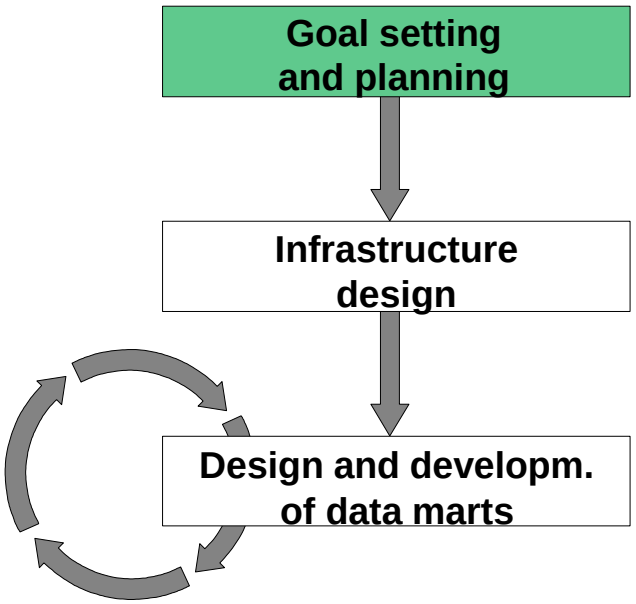


Top-down:
1. Design of DW
2. Design of DMs

"In-between":
1. Design of DW for DM1
2. Design of DM2 and integration with DW
3. Design of DM3 and integration with DW
4. ...

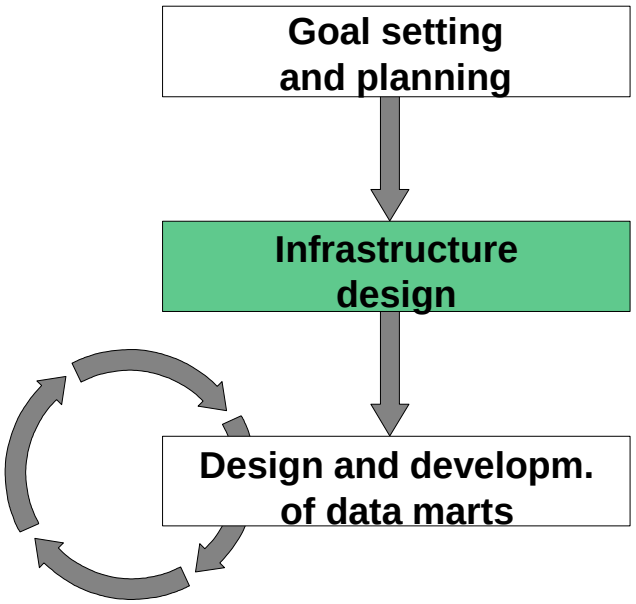
Bottom-up:
1. Design of DMs
2. Integration of DMs in DW
3. Maybe no physical DW

The Life-cycle/ I



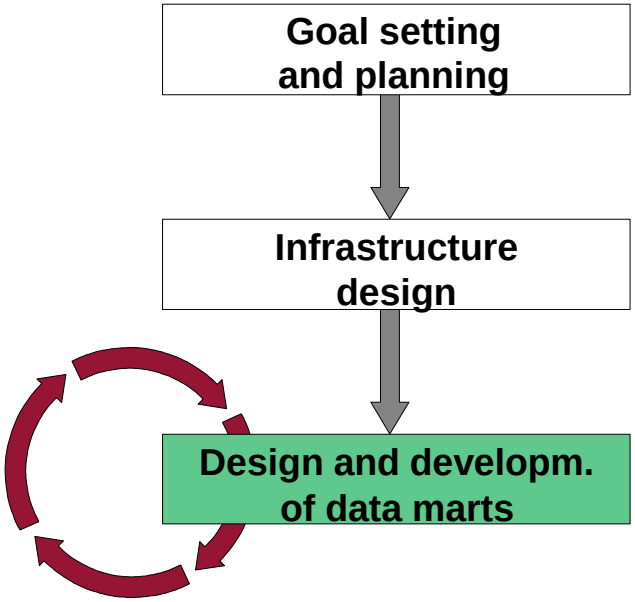
- set system goals, borders, and size
- select an approach for design and implementation
- estimate costs and benefits
- analyze risks and expectations
- examine the skills of the working team

The Life-cycle/2



- analyze and compare the possible architectural solutions
- assess the available technologies and tools
- create a preliminary plan of the whole system

The Life-cycle/3



Every iteration causes a new data mart and new applications to be created and progressively added to the DW system

Data mart design phases

Source analysis and integration

db administrator

Requirement analysis

Conceptual design

business user

Workload and data volume

designer

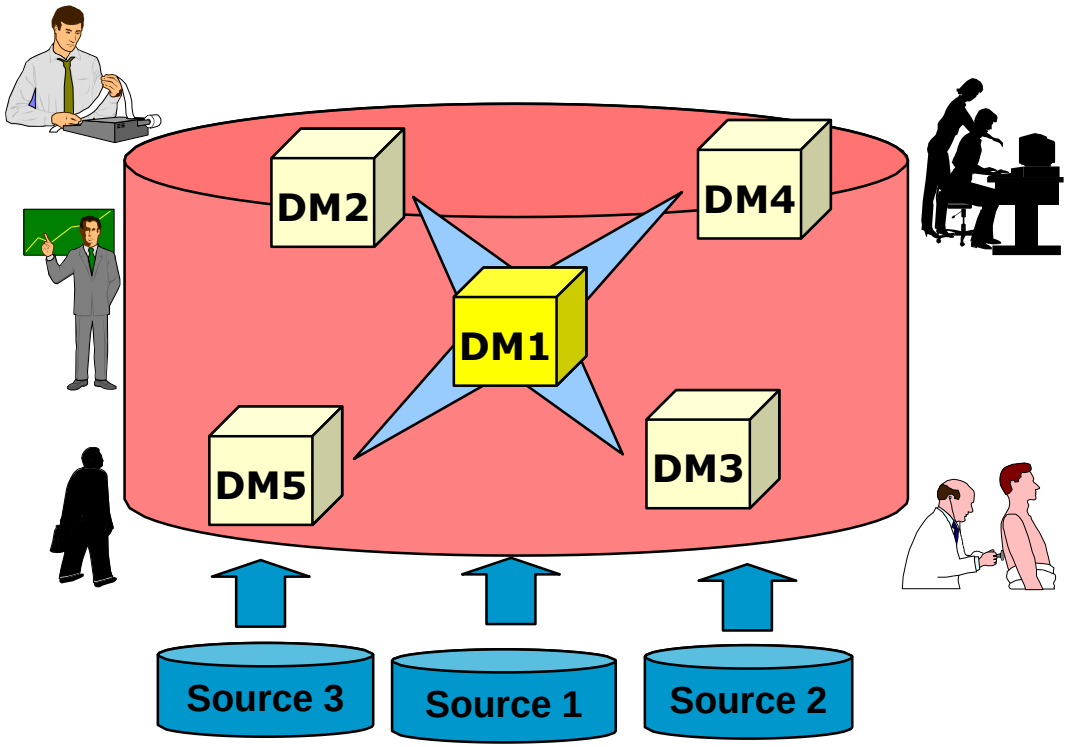
Logical design

ETL design

Physical design

The First Data Mart

- is the one playing the most strategic role for the enterprise
- should be a backbone for the whole DW
- should lean on available and consistent data sources



Summary

- BI is well-recognized and is a combination of a number of techniques to support decision making.
- DW is at the core of BI that
  - provides a complete, consistent, subject-oriented and time-varying collection of the data;
  - allows to separte OLTP from OLAP.
- Applications that **use** the DW include OLAP, data mining, visualization
- BI can provide many advantages to an organization
  - Creates added value by transforming data into information
  - Provides comprehensive knowledge about your business
  - 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
- Following a clear design methodology is important.