Information Integration

Part 1: Basics of Relational Database Theory

Werner Nutt

Faculty of Computer Science
Master of Science in Computer Science

A.Y. 2011/2012

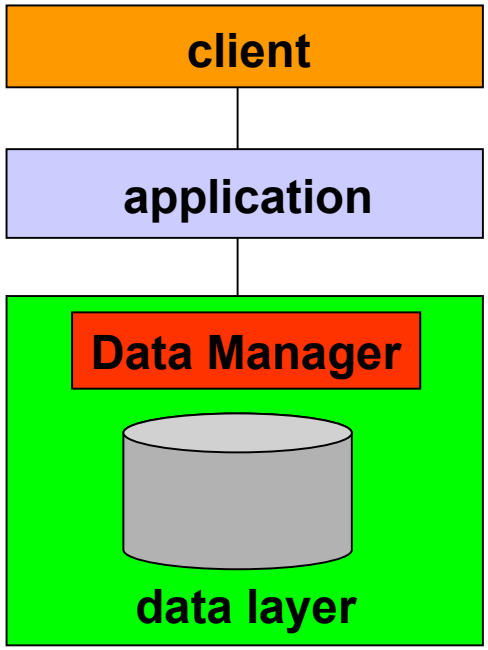
Freie Universität Bozen
Libera Università di Bolzano
Free University of Bozen · Bolzano

Integration in Data Management: Evolution

- The Classical Database Application
- Database Application with Several DBMSs
- Data Access via Distributed DBMS
- Federated Database System
- Data Integration (with Global Schema)

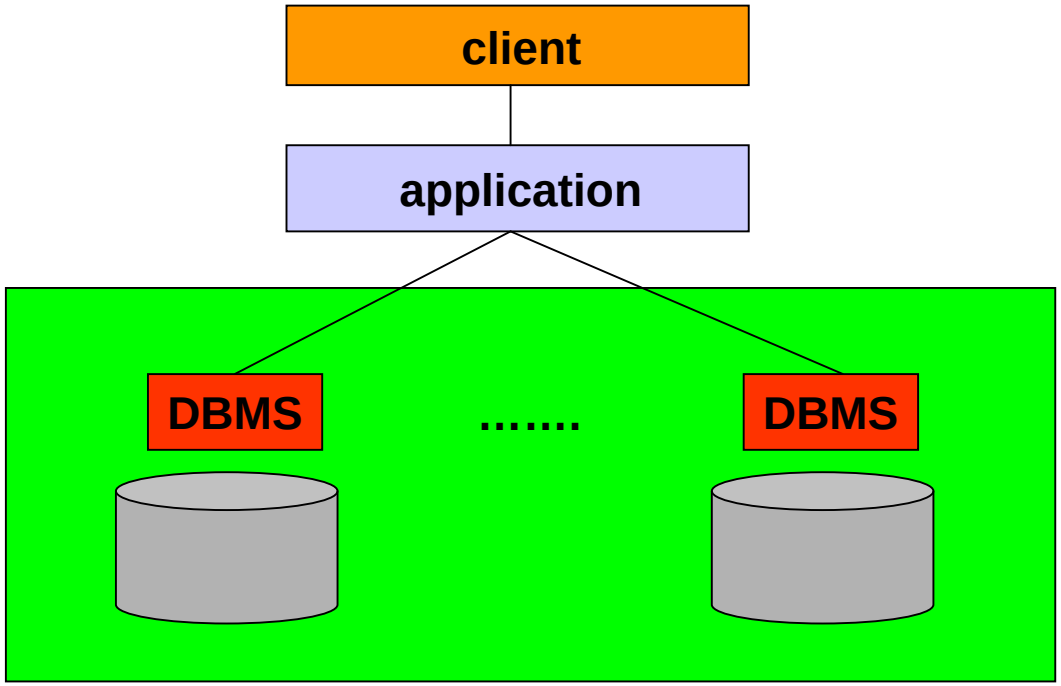
Slides on evolution due to Maurizio Lenzerini

The Classical Database Application



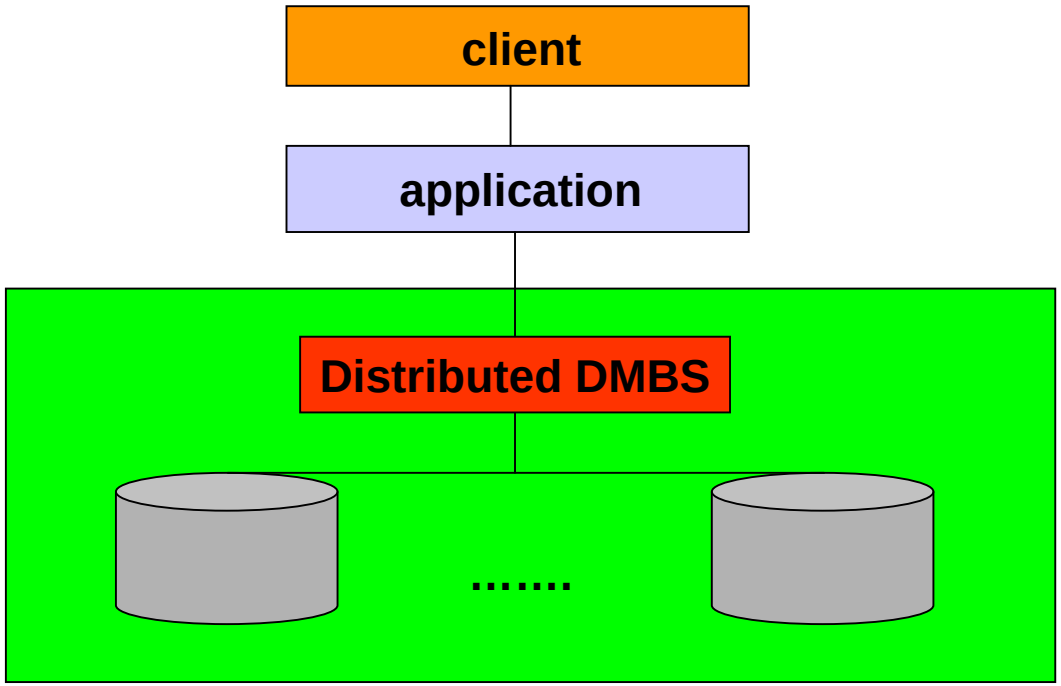
- Centralized system with three-tier architecture
- Implicit integration: integration supported by the Data Base Management System (DBMS), i.e., the data manager

Database Application with Several DBMSs



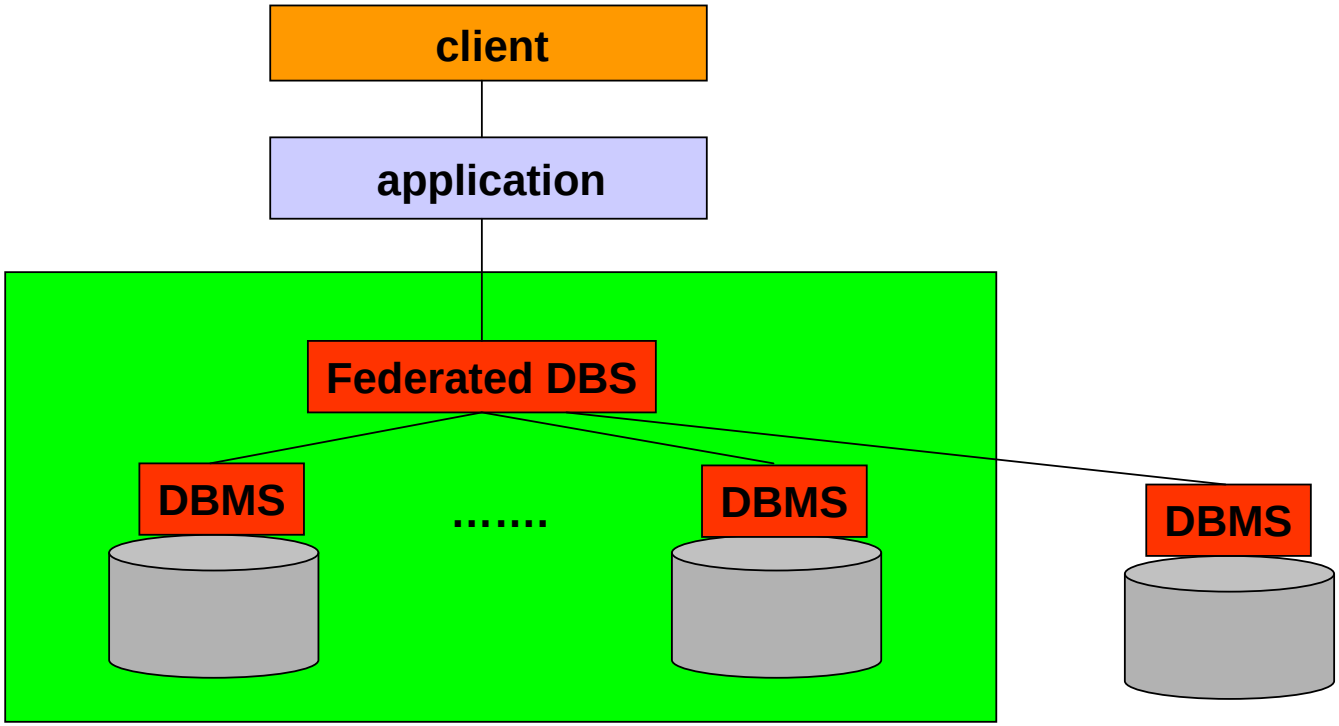
- Centralized system with three-tier architecture and multiple stores
- Application hides integration: integration "embedded" within application

Data Access via Distributed DBMS



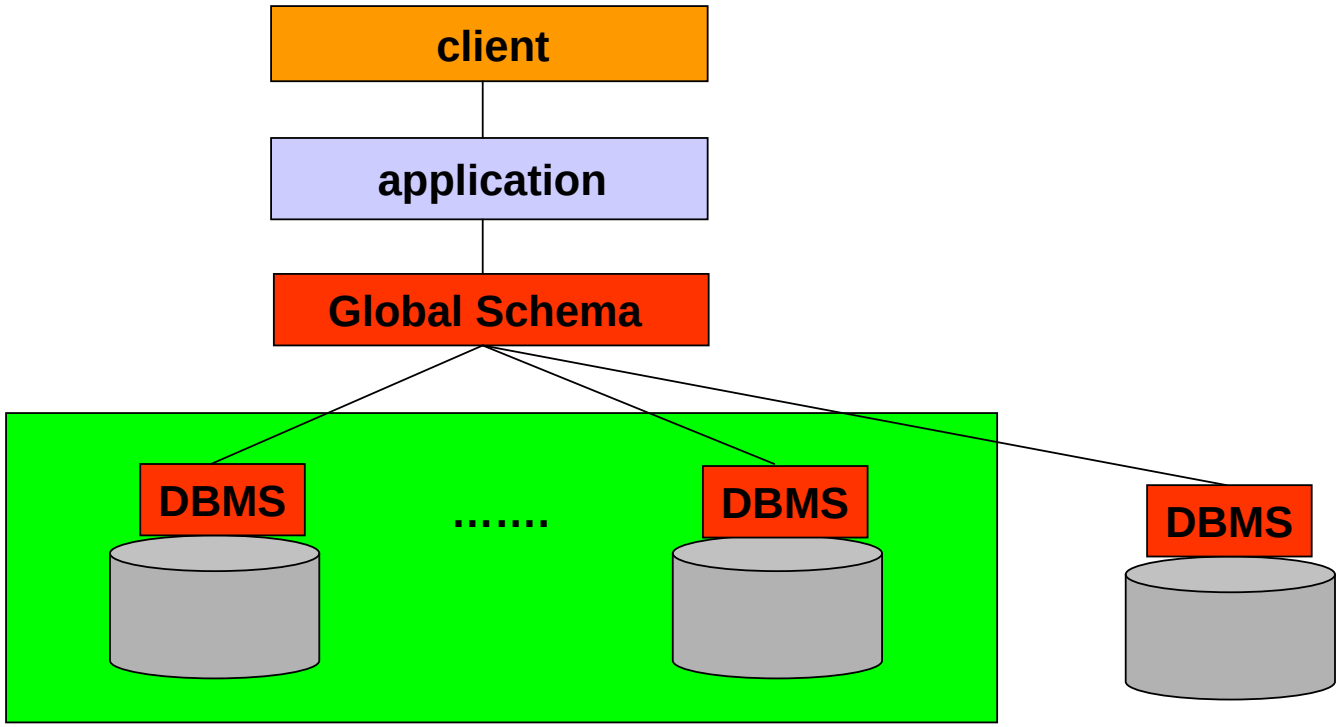
- Centralized system with three-tier architecture and multiple data stores
- Distributed data management: different data sources of the same type, under the control of the organization, managed by a Distributed DBMS

Federated Database System



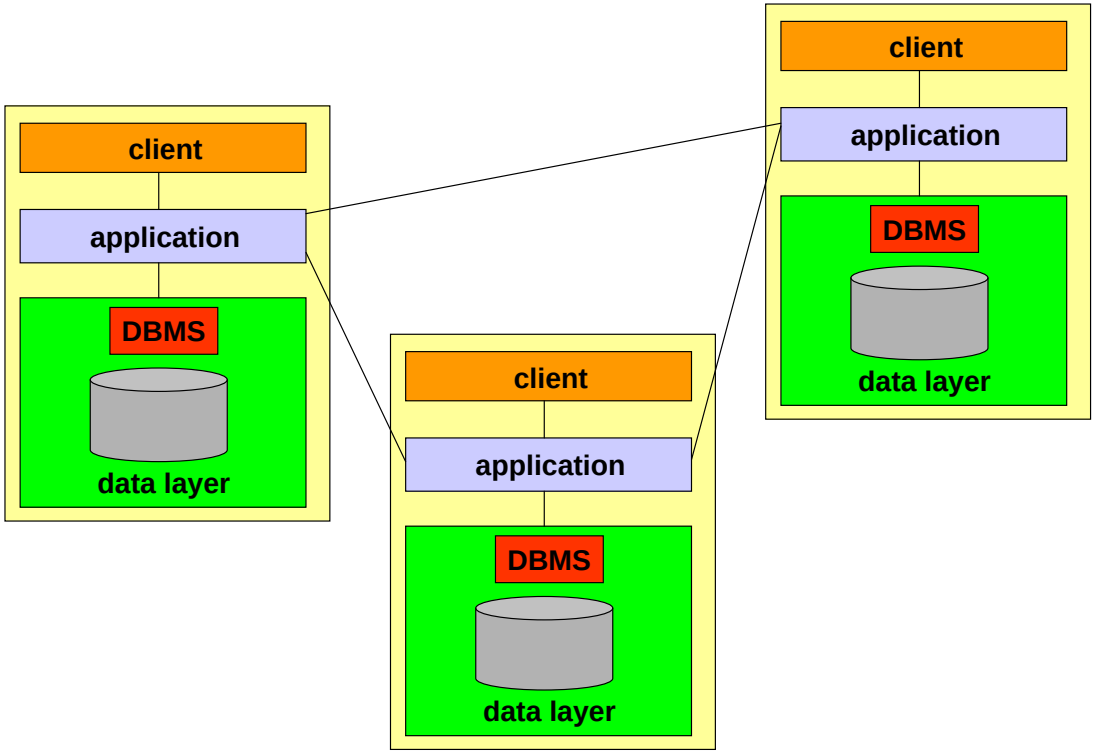
- Centralized system with three-tier architecture and distributed stores
- Data federation: different data sources, not necessarily of the same type, or under the control of the organization, federated within one data layer

Data Integration (with Global Schema)



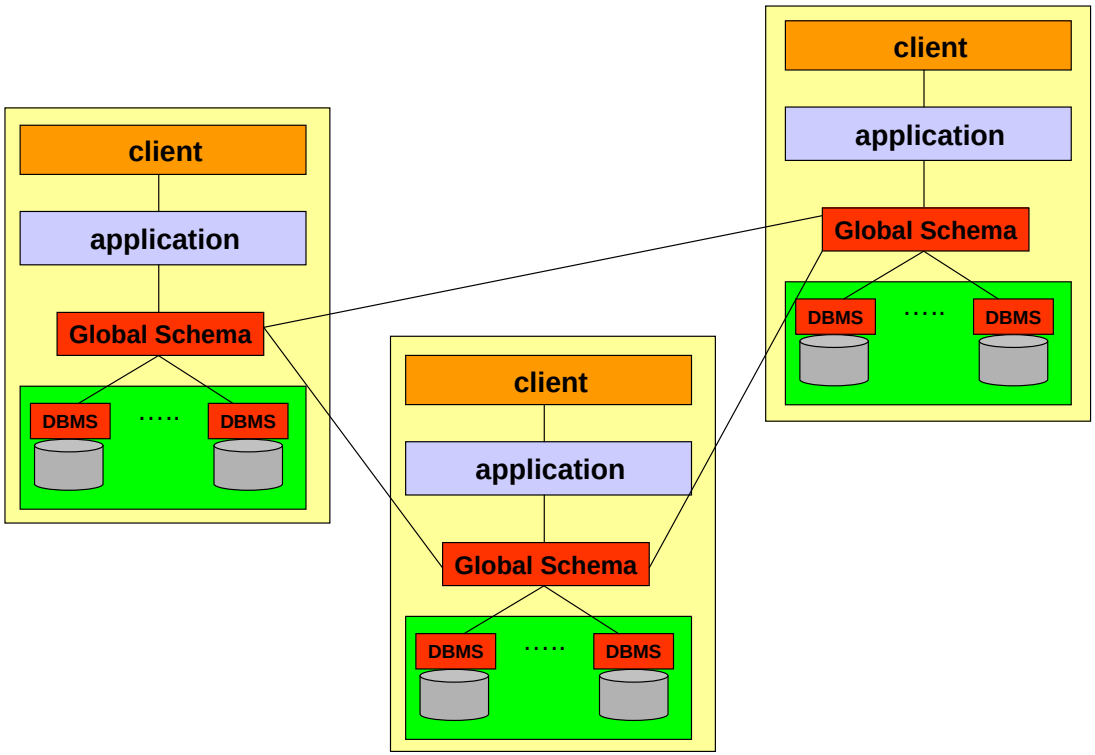
- Centralized system with four-tier architecture and distributed stores
- Data exchange and integration: the global schema is "independent" from the different data sources, which are heterogeneous, and not necessarily under the control of a single organization

Application-based Distribution



- Decentralized system
- Application-based distribution: distributed integration realized within application

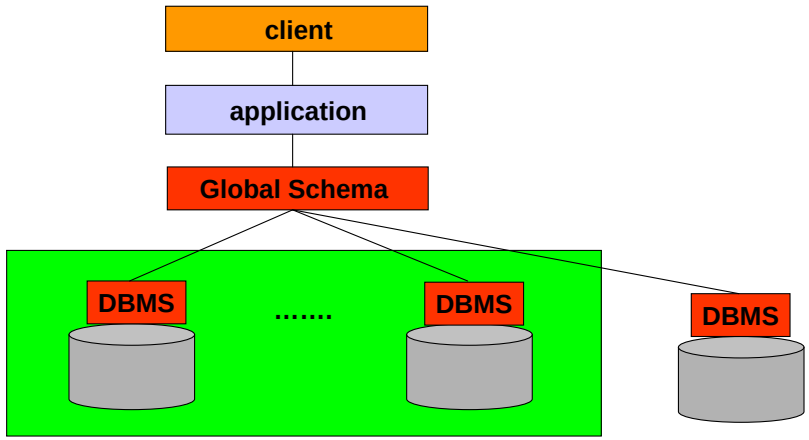
P2P Data Integration



- Centralized system with three-tier architecture
- Peer-to-peer data exchange and integration: distributed data integration realized with no central global schemas

What is Information Integration?

- Information Integration is the problem of
  - providing a unified and transparent view
  - to a collection of data stored in multiple, autonomous, and heterogeneous data sources.
- The unified view is achieved through a global (or target) schema, and is realized either
  - through a materialized database (exchange), or
  - through a virtualization mechanism based on querying (integration).



Relevance of Information Integration

- Growing demand (and market)
- One of the major challenges for the future of IT
- At least two contexts
  - Intra-organization information integration (e.g., EIS)
  - Inter-organization information integration (e.g., integration on the Web)

Information Integration: Available Industrial Solutions

- Distributed database systems
- Tools for source wrapping
- Tools for ETL (Extraction, Transformation and Loading)
- Data warehousing
- Tools based on database federation, e.g., DB2 Information Integrator
- Distributed query optimization

Current Information Integration Tools: Characteristics

- Physical transparency, i.e., masking from the user the physical characteristics of the sources
- Heterogeinity, i.e., federating highly diverse types of sources
- Extensibility
- Autonomy of data sources
- Performance, through distributed query optimization

However, current tools do not (directly) support the so-called logical (or conceptual) transparency (via an integrated schema), which is crucial in data integration

Theme of This Course

- Databases are everywhere these days
- Every enterprise has a database; they merge, combine data hence data integration
- In addition, a lot of data is available on the web, but often one needs many sources to answer a query
- Hence (almost) everyone needs to integrate data
- Huge investment from leading companies, IBM, Oracle, Microsoft
- Very ad hoc solutions; but finally we understand what the real problems in data integration are, and have some solutions (but not all!)

Objectives of the Course

- Introduce the formal concepts from the area of databases by which information integration problems are modeled
  I.e., the concepts you find in research papers
- Present techniques for
  - Mapping schemas to each other
  - Evaluating queries in this setting
  - Assessing the quality of query answers
- Train fundamental mathematical skills such as
  - giving formal definitions
  - formulating theorems
  - proving or disproving formal statements.

Topics

- Basics of Relational Database Theory
- Modeling Information Sources: Global as View, Local as View
- Query Semantics and Query Planning
- Data Quality: Consistency and Completeness
- Possibly a glimpse at:
  - Data Exchange
  - Sources with Access Limitations (Forms, Web Services)
  - Constructing Schema Mappings

Course Organisation

- No textbook, since none exists (but survey and research papers)
- Slides, papers, and links to further info will be posted on course website (reachable from my home page)
- Coursework:
  - 5 sets of exercises (up to 30% of total mark, depending on the number of correctly solved exercises)
  - possibly, a presentation on an information integration tool (20% of total mark)
- Coursework mark plays two roles
  - for passing, the pass mark has to be at least 18:
    pass mark $=$ max {exam mark, $0.7 \times$ exam mark $+ 0.3 \times$ exercise mark}
  - the final mark can be improved by a presentation (see above)
    final mark $=$ max {exam mark, $0.7 \times$ exam mark $+ 0.3 \times$ exercise mark $+$ $0.2 \times$ presentation}