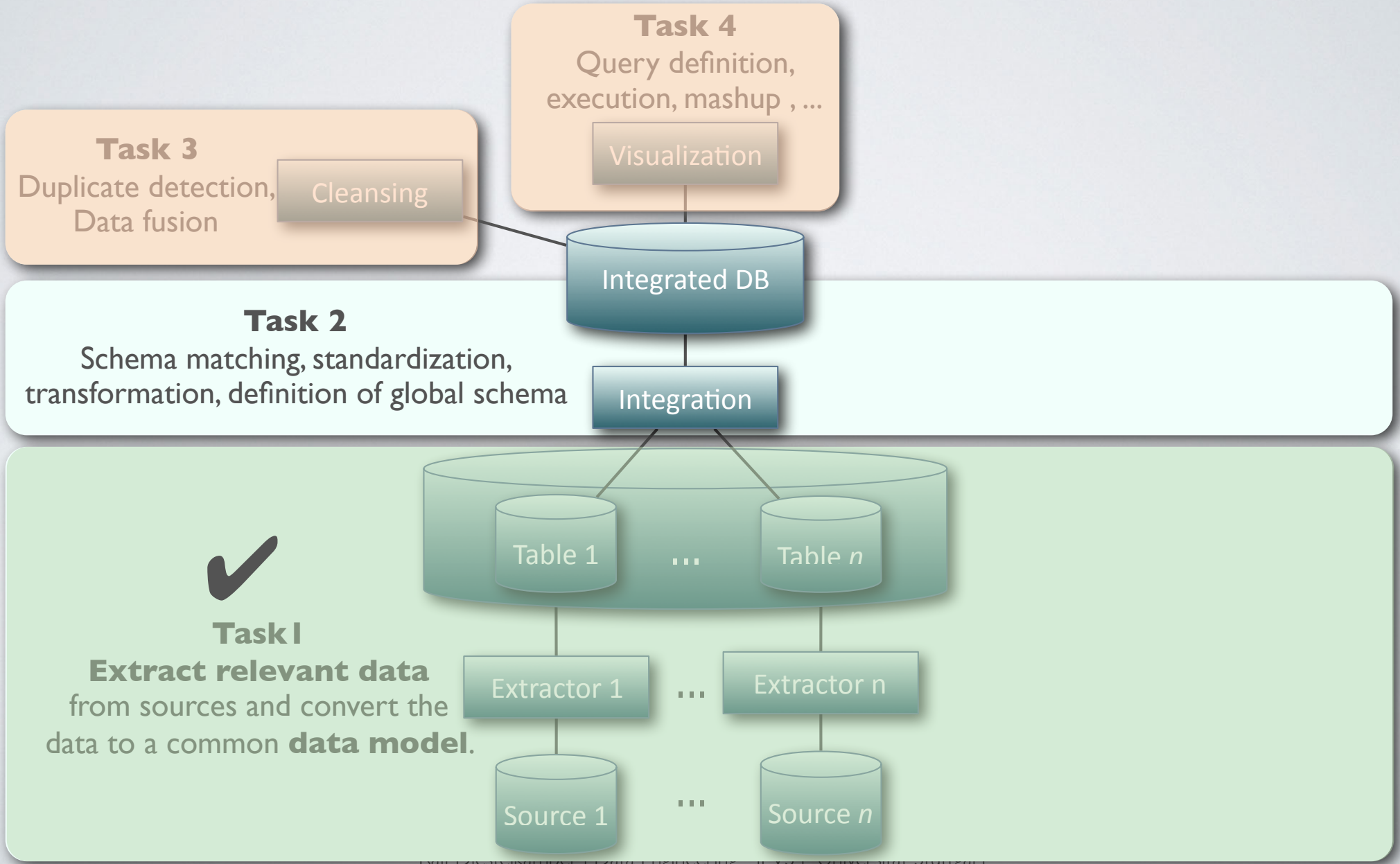
# INFORMATION INTEGRATION

Practicals Winter Term 2016/17

Ralf Diestelkämper | Data Engineering - IPVS | Universität Stuttgart

# STATUS


Task 4
Query definition, execution, mashup , ...
Visualization
Task 3
Duplicate detection, Data fusion
Cleansing
Integrated DB
Task 2
Schema matching, standardization, transformation, definition of global schema
Integration
Table 1
...
Table n
Task I
Extract relevant data
from sources and convert the data to a common data model.
Extractor 1
...
Extractor n
Source 1
...
Source n

# WEEKLY SCHEDULE

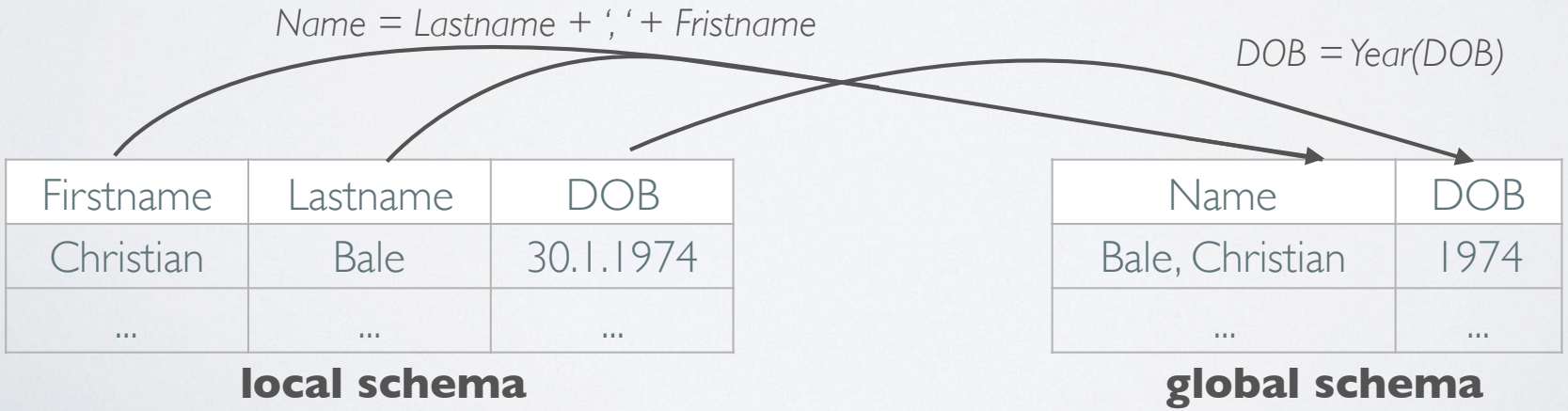
| Practicals | Practical | Project | Handout |
|---|---|---|---|
| 26.10.2016 | | Practicals & project overview<br>Introduction Project Task 1 | Exercise sheet 1 |
| 2.11.2016 | Correction sheet 1 | | Exercise sheet 2 |
| 9.11.2016 | | Project presentations Task 1 +<br>Introduction Project Task 2 | |
| 16.11.2016 | Correction sheet 2 | | Exercise sheet 3 |
| 23.11.2016 | Correction sheet 3 | | Exercise sheet 4 |
| 30.11.2016 | | Project presentations Task 2 | |
| 7.12.2016 | Correction sheet 4 | | Exercise sheet 5 |

# TASK 2: INTEGRATION

- Define a **global schema**
  - Any data model is acceptable (e.g., relational, hierarchical, object oriented)
  - Avoid redundancy
    - in the schema (e.g., each entity type & attribute appears only once)
    - in the data (normalization)
  - Only keep information relevant to your application (If not already filtered during prior extraction phase).

# TASK 2: INTEGRATION

- Define a **schema matching** between the source schemas and the global (target) schema
  - Automatically if possible(easy solutions are OK!)
  - Manually as fallback if automatic matching not accurate enough.



# TASK 2: USEFUL LINKS

- Helpful **links** for **automatic schema matching**
  - Available for free: COMA++
    http://dbs.uni-leipzig.de/Research/coma.html
  - Function Library for your own implementation: SecondString
    http://secondstring.sourceforge.net/
  - Algorithms for automatic schema matching: see lecture

# TASK 2: STANDARDIZATION

- **Standardization of the source data**
  - All values of a same (global) attribute should have the same format
  - Examples:
    - American (MM/DD/YYYY) vs German (DD.MM.YYY) data format
    - Names: "Brad Pitt" vs. "Pitt, Brad" vs. "Mr. Pitt" vs. "B. Pitt" ...
    - Early standardization simplifies subsequent data cleaning task (with the focus on entity resolution).

# TASK 2: TRANSFORMATION

- **Transforming** the source data into the target representation
    - For **materialized** integration: load data in your target schema
    - For **virtual** integration: Write wrappers to fill all data necessary in your global schema.
    - Perform **standardization** at this stage.
    - Consider the **schema matching** you determined.

# TASK 2: PRESENTATIONS

- Your presentation should  at least include
  - A detailed description of your global (integrated) schema
  - Your schema matching and a description of your automated schema matching solution.
  - The standardization that was necessary with actual examples.
  - Examples of your source data and the transformed target data illustrating the functionality of your integration.
  - A "proof" that your application requires the integrated data, i.e., a single source cannot answer your application query.

# TASK 2: PRESENTATIONS

- Create slides to present your solution
- Put your slides into ILIAS
  - The day before your presentation at 3 p.m the latest
  - As PDF, PPT(X), Keynote or Open Office file)
- Language: English
- Date: 20.11.2015
- Duration: 7 - 10 min (depending on final number of groups)
- Presence is mandatory
  - -0.3 on internal grade translated to passing or failing the project

# TASK2: PRESENTATIONS

- Create slides to present your solution
- Put your slides into ILIAS
  - The day before your presentation at 3 p.m the latest
  - As PDF, PPT(X), Keynote or Open Office file)
- Language: English
- Date: 30.11.2016
- Duration: 7 - 10 min
- Presence is mandatory