HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Information Integration
Course Project

Thorsten Papenbrock

WS 2015 / 2016

Database Systems II
Übung Tutor

Information Integration Project

ThorstenPapenbrock,
WS 2015 / 2016

Information Integration
Course Project – Organization

1. Big data integration project parallel to the lecture with all participants

2. Teams of 4 Students

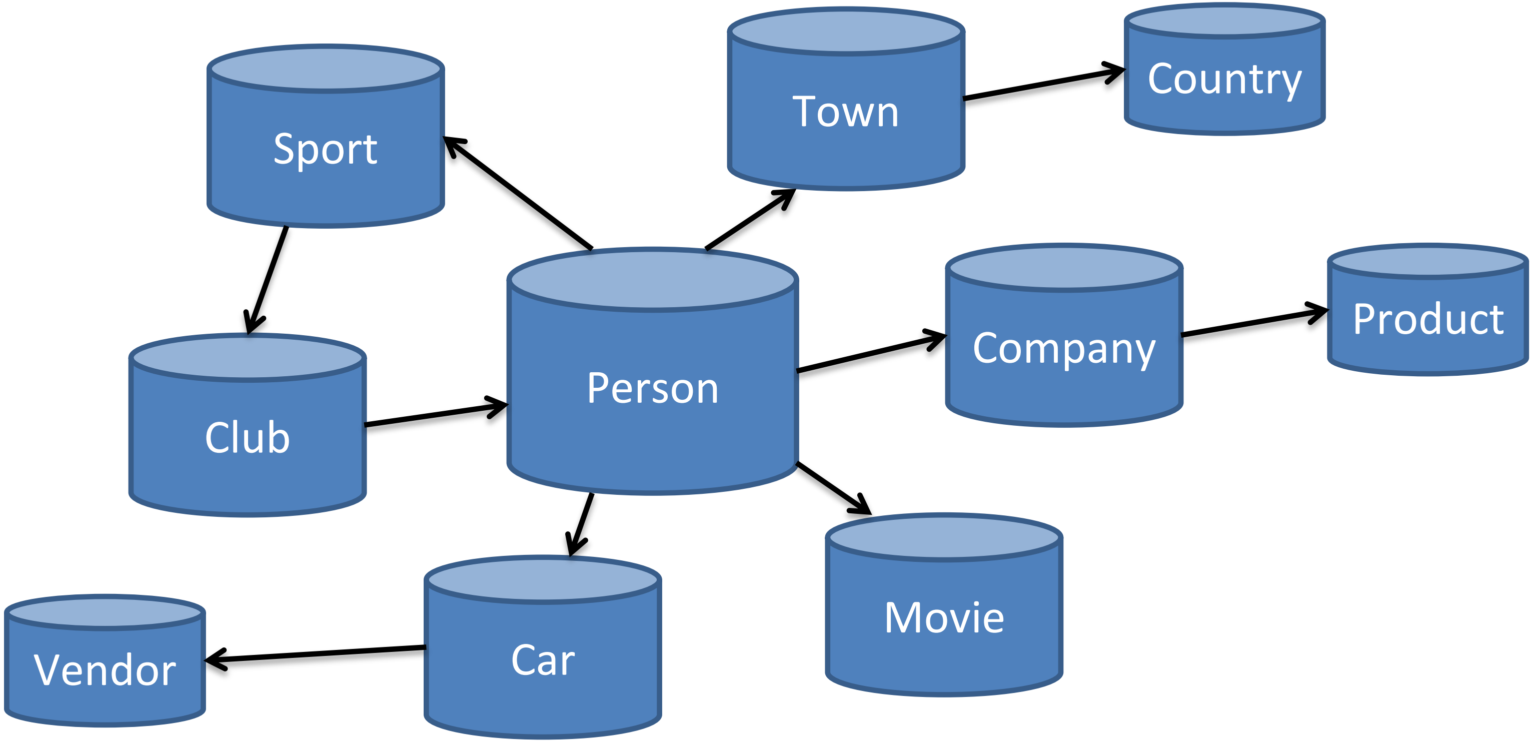
3. Presentation of sub-task results in the exercise lectures

4. Grading of sub-tasks and presentations: excellent, good, failed

5. Successfully passing the project is important for exam permission

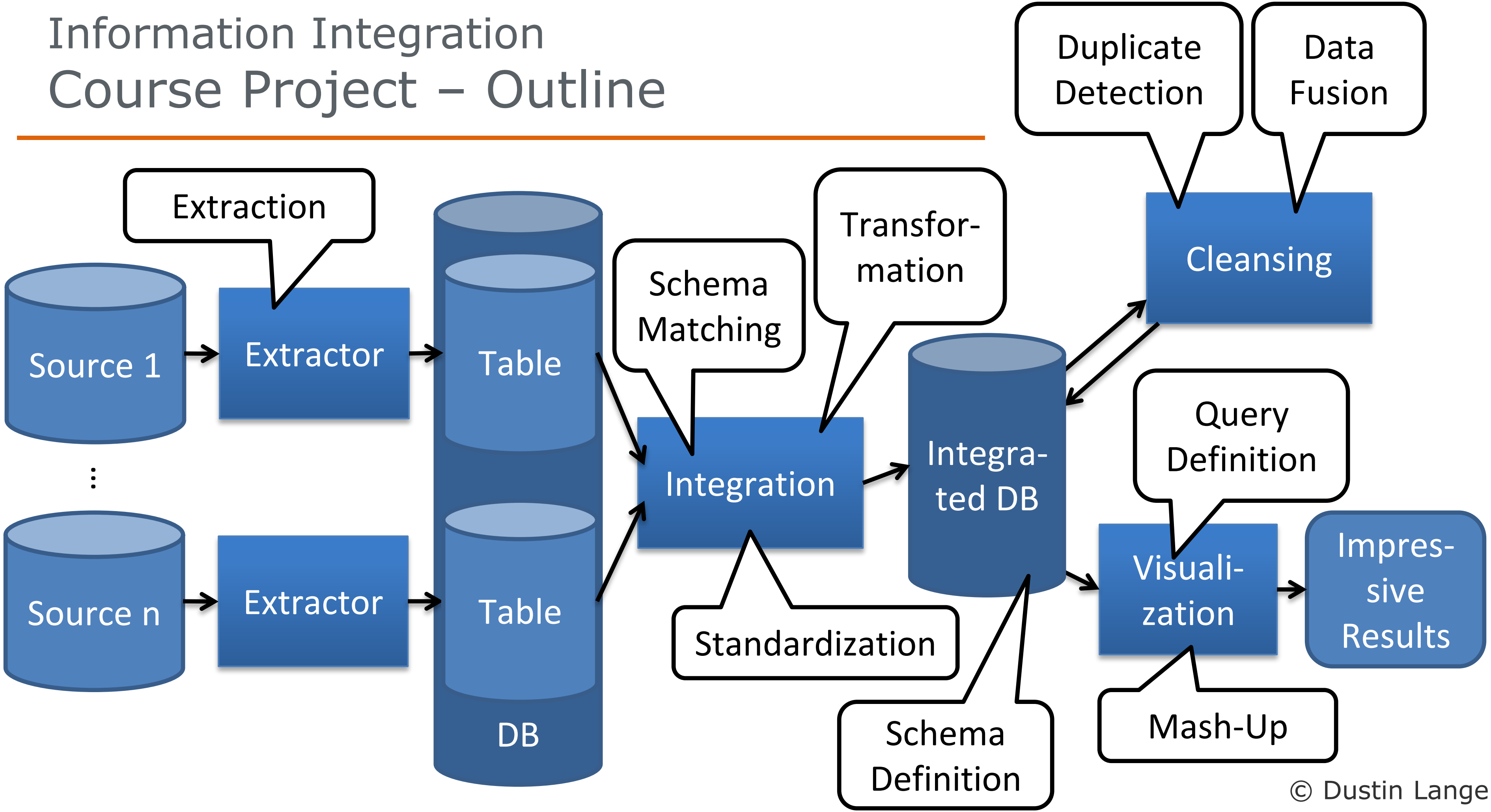
Information Integration
Course Project – Vision

Integrated database on public personalities



Information Integration

Course Project – Outline



© Dustin Lange

Information Integration
Course Project – Tasks

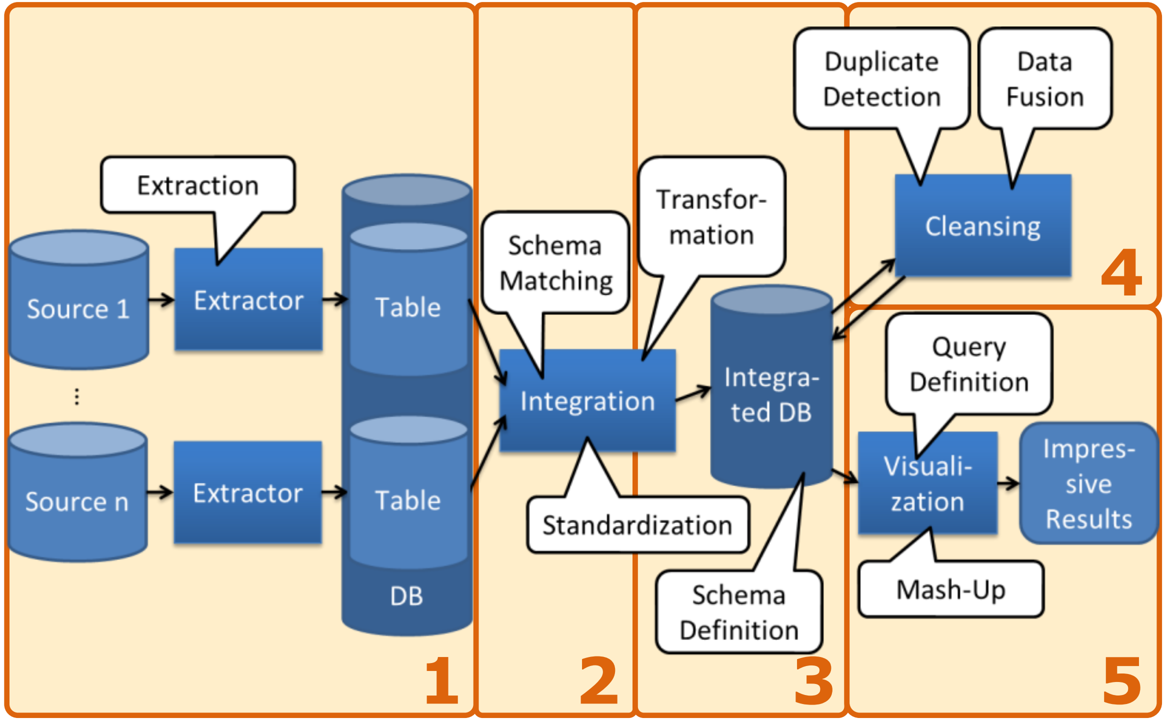
Task 1: Extraction

Task 2: Integration Planning

Task 3: Integration Execution

Task 4: Cleansing

Task 5: Visualization



Information Integration
Course Project – Deliverables

- **3-5 Slides**
  - for <5 min presentations in class
  - showing your ideas, techniques, and results (in German or English)
  - in pdf format
  - with name <last-name1>_ <last-name2>_ <last-name3>_ <last-name4>.<pdf>

- *Submission:*
  - *Channel: Email at thorsten.papenbrock(at)hpi.de*
  - *Subject: [InfoInt2015] Exercise <NR> <last-name1>*
  - *Deadline: Two work days before exercise lectures*
    → Monday for Wednesday lectures

Note: Do not forget the author names on your slides! →

Information Integration
Course Project – Timetable

Title	Date	Periode	Introduction	Presentation
Exercise 1	21.10.15	3 weeks	Extraction	
Exercise 2	11.11.15	2 weeks	Integration Planning	Extraction
Exercise 3	25.11.15	3 weeks	Integration Execution	Integration Planning
Exercise 4	16.12.15	3 weeks	Cleansing	Integration Execution
Exercise 5	13.01.16	3 weeks	Visualization	Cleansing
Exercise 6	03.02.16			Visualization

Information Integration
Course Project – Extraction

1. Find 2 datasets that each …
   - contain public person data (politicians, athletes, actors, …).
   - contain at least 1 and at most 4 additional entities (party, sport, movie, …).
   - contain more than 4 attributes.
   - originate from different sources / web sites.

2. Design a database schema for each dataset individually that …
   - captures the data as it is (no standardization and no cleansing!).

3. Extract the data you have chosen and insert it into your new schema …
   - using a self written extractor or a tool that you found in the internet.

4. Document your schemata in your presentation slides and …
   - introduce the source datasets (topic, size, source, …).
   - provide the ER-diagram.
   - provide the create table statements (fields, datatypes).
   - provide the add constrain statements (keys, foreign keys).

Information Integration
Course Project – Dataset Examples

Persondata on Wikipedia

- Example:
```
{{Persondata
  | NAME                = Gandhi, Mohandas Karamchand
  | ALTERNATIVE NAMES   = Gandhi, Mahatma
  | SHORT DESCRIPTION   = Political leader
  | DATE OF BIRTH       = 2 October 1869
  | PLACE OF BIRTH      = Porbandar, Gujarat, India
  | DATE OF DEATH       = 30 January 1948
  | PLACE OF DEATH      = Birla House, New Delhi, India
}}
```
- Link:
  - https://en.wikipedia.org/wiki/Wikipedia:Persondata

Information Integration
Course Project – Dataset Examples

Persondata on DBpedia

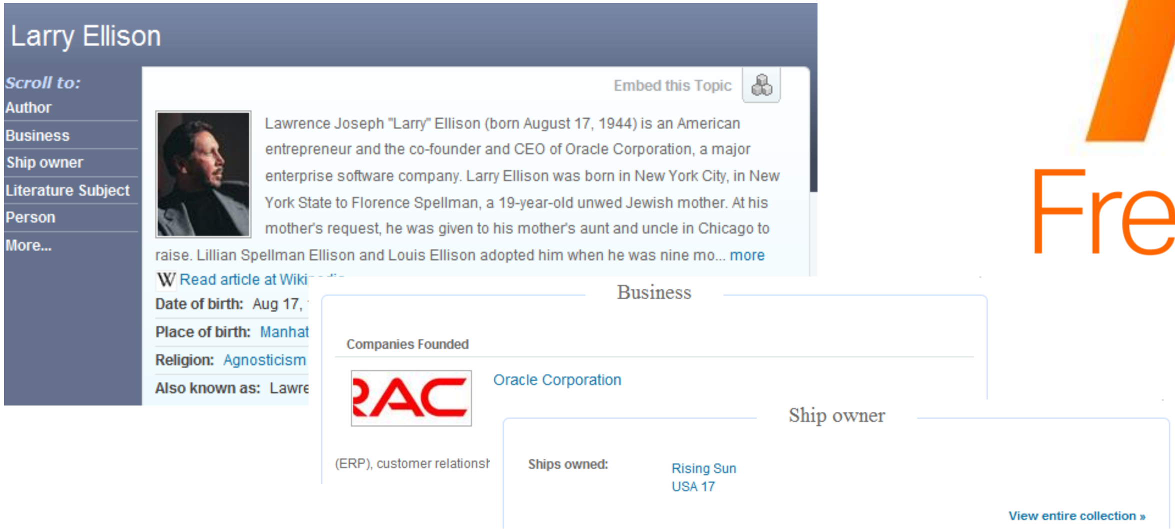
- Example:
```
<http://viaf.org/viaf/71391324>
         <rdf:type> <schema:Person>
         <schema:birthDate> "1869”
         <schema:deathDate> "1948”
         <schema:name> "Gandhi, Mahatma"
```

- Link:
  - http://wiki.dbpedia.org/Downloads2015-04#persondata

Information Integration
Course Project – Dataset Examples

Data on Freebase

- Incorporates data from: Wikimedia, MusicBrainz, WordNet, …
- Example:



Larry Ellison

Scroll to:
Author
Business
Ship owner
Literature Subject
Person
More...

Embed this Topic

Lawrence Joseph "Larry" Ellison (born August 17, 1944) is an American entrepreneur and the co-founder and CEO of Oracle Corporation, a major enterprise software company. Larry Ellison was born in New York City, in New York State to Florence Spellman, a 19-year-old unwed Jewish mother. At his mother's request, he was given to his mother's aunt and uncle in Chicago to raise. Lillian Spellman Ellison and Louis Ellison adopted him when he was nine mo... more

W Read article at Wiki

Date of birth: Aug 17,

Place of birth: Manhat

Religion: Agnosticism

Also known as: Lawre

Business

Companies Founded

Oracle Corporation

(ERP), customer relationsh

Ship owner

Ships owned: Rising Sun
USA 17

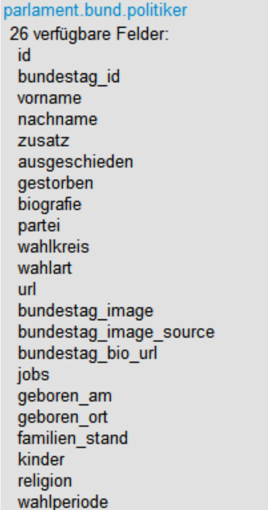
View entire collection »

- Link:
  - http://wiki.freebase.com/wiki/Main_Page

Information Integration
Course Project – Dataset Examples

Data from DeutschlandAPI

- Example:
  ```
  { "id":"6769",
    "titel":"Verbraucherschutz - Umweltampel ",
    "beschreibung":null,
    "text":null,
    "hauptpetent":"60",
    "status":"in der Mitzeichnung",
    "bundestag_board_id":"1352.0",
    "ended":"2009-11-04 01:00:00",
    "started":"2009-08-19 02:00:00",
    "system_updated":"2009-10-29 17:40:41„ }
  ```
- Link:
  - http://www.deutschland-api.de/Api



parlament.bund.politiker

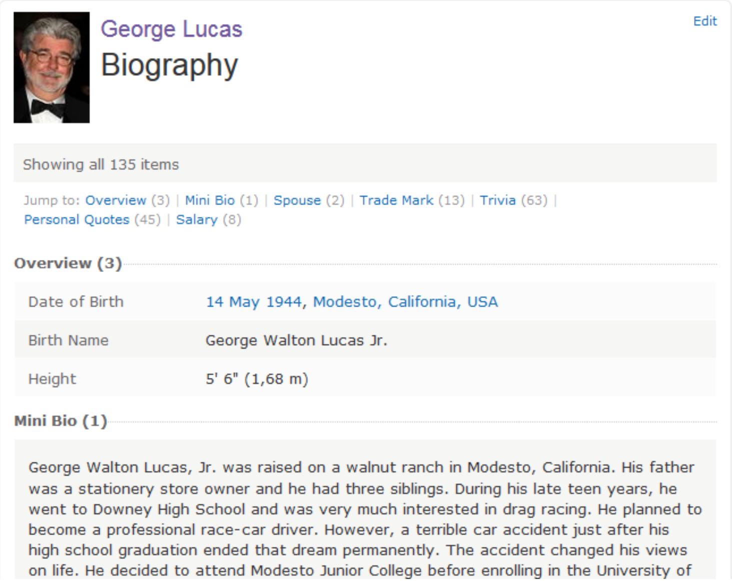
26 verfügbare Felder:

id
bundestag_id
vorname
nachname
zusatz
ausgeschieden
gestorben
biografie
partei
wahlkreis
wahlart
url
bundestag_image
bundestag_image_source
bundestag_bio_url
jobs
geboren_am
geboren_ort
familien_stand
kinder
religion
wahlperiode

Information Integration
Course Project – Dataset Examples

Data from IMDB

- Example:



George Lucas
Biography

Edit

Showing all 135 items

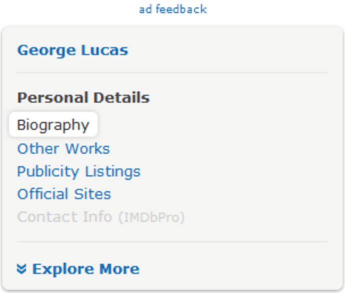
Jump to: Overview (3) | Mini Bio (1) | Spouse (2) | Trade Mark (13) | Trivia (63) | Personal Quotes (45) | Salary (8)

Overview (3)

Date of Birth	14 May 1944, Modesto, California, USA
Birth Name	George Walton Lucas Jr.
Height	5' 6" (1,68 m)

Mini Bio (1)

George Walton Lucas, Jr. was raised on a walnut ranch in Modesto, California. His father was a stationery store owner and he had three siblings. During his late teen years, he went to Downey High School and was very much interested in drag racing. He planned to become a professional race-car driver. However, a terrible car accident just after his high school graduation ended that dream permanently. The accident changed his views on life. He decided to attend Modesto Junior College before enrolling in the University of



ad feedback

George Lucas

Personal Details

Biography
Other Works
Publicity Listings
Official Sites
Contact Info (IMDbPro)

Explore More

- Link:
  - http://www.imdb.com/interfaces

Information Integration
Course Project – Dataset Examples

Data about baseball players

- Schema:

birthYear	Year player was born
birthMonth	Month player was born
birthDay	Day player was born
birthCountry	Country where player was born
birthState	State where player was born
birthCity	City where player was born
deathYear	Year player died
deathMonth	Month player died
deathDay	Day player died
deathCountry	Country where player died
deathState	State where player died
deathCity	City where player died
nameFirst	Player's first name
nameLast	Player's last name
nameNote	Note about player's name
nameGiven	Player's given name (typically first and middle)
nameNick	Player's nickname
weight	Player's weight in pounds
height	Player's height in inches
bats	Player's batting hand (left, right, or both)
throws	Player's throwing hand (left or right)

- Link:
  - http://seanlahman.com/baseball-archive/statistics/

SeanLahman.com

Information Integration
Course Project – Dataset Examples

NO GENERATED DATA

Information Integration
Course Project – Coordination

Dataset selection and proposals:

http://doodle.com/poll/7yp44t5zci3g2pr2

- Add as many datasets as you find with this pattern:
  <Dataset Name> (<URL>)
- Select only those datasets that you take for your team.
- Each dataset should be selected by only one team!

Information Integration
Course Project – Extraction

1. Find 2 datasets that each …
   - contain public person data (politicians, athletes, actors, …).
   - contain at least 1 and at most 4 additional entities (party, sport, movie, …).
   - contain more than 4 attributes.
   - originate from different sources / web sites.

2. Design a database schema for each dataset individually that …
   - captures the data as it is (no standardization and no cleansing!).

3. Extract the data you have chosen and insert it into your new schema …
   - using a self written extractor or a tool that you found in the internet.

4. Document your schemata in your presentation slides and …
   - introduce the source datasets (topic, size, source, …).
   - provide the ER-diagram.
   - provide the create table statements (fields, datatypes).
   - provide the add constrain statements (keys, foreign keys).

Information Integration
Course Project – Extraction

- Hints:
  - Most data sources also provide their *data types*.
  - Many data sources also provide *relational schemata*.
  - Some RDF, Json and XML data sources also provide *relational parser*.
  - Pay attention to the data *encoding* of your datasets!
  - Do not forget to define keys and foreign-keys!

Information Integration
Course Project – Extraction

1. Find 2 datasets that each …
   - contain public person data (politicians, athletes, actors, …).
   - contain at least 1 and at most 4 additional entities (party, sport, movie, …).
   - contain more than 4 attributes.
   - originate from different sources / web sites.

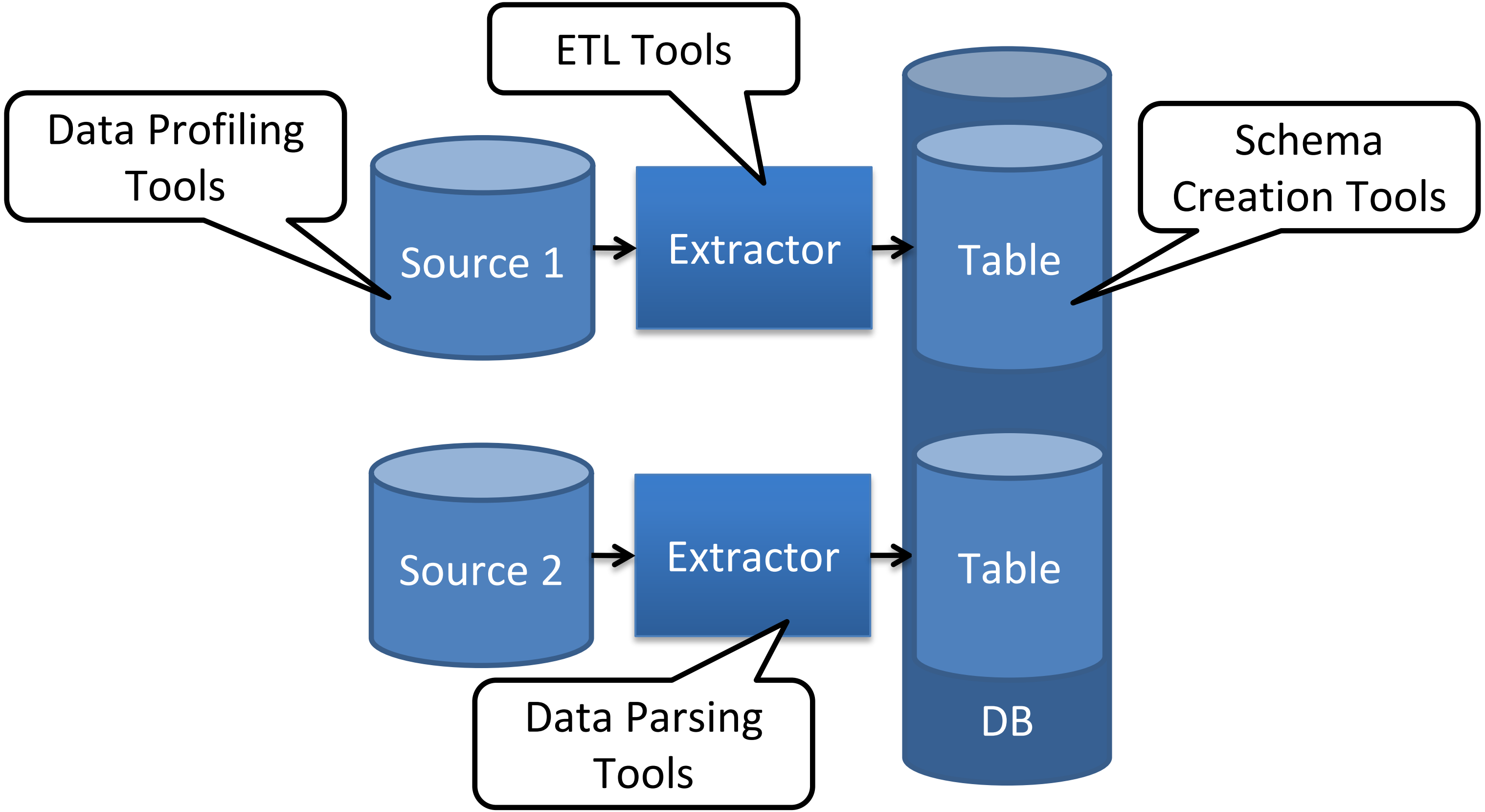
2. Design a database schema for each dataset individually that …
   - captures the data as it is (no standardization and no cleansing!).

3. Extract the data you have chosen and insert it into your new schema …
   - using a self written extractor or a tool that you found in the internet.

4. Document your schemata in your presentation slides and …
   - introduce the source datasets (topic, size, source, …).
   - provide the ER-diagram.
   - provide the create table statements (fields, datatypes).
   - provide the add constrain statements (keys, foreign keys).

Information Integration
Course Project – Database

Information Integration
Course Project – Using Tools?



Information Integration
Course Project – Profiling (and ETL) Tools

IBM InfoSphere Information Analyzer
http://www.ibm.com/software/data/infosphere/information-analyzer/

Oracle Enterprise Data Quality
http://www.oracle.com/us/products/middleware/data-integration/enterprise-dataquality/overview/index.html

Talend Data Quality
http://www.talend.com/products/data-quality

Ataccama DQ Analyzer
http://www.ataccama.com/en/products/dq-analyzer.html

SAP BusinessObjects Data Insight
http://www.sap.com/germany/solutions/sapbusinessobjects/large/eim/datainsight/index.epx

SAP BusinessObjects Information Steward
http://www.sap.com/germany/solutions/sapbusinessobjects/large/eim/information-steward/index.epx

Informatica Data Explorer
http://www.informatica.com/us/products/data-quality/data-explorer/

Microsoft SQL Server Integration Services Data Profiling Task and Viewer
http://msdn.microsoft.com/en-us/library/bb895310.aspx

Trillium Software Data Profiling
http://www.trilliumsoftware.com/home/products/data-profiling.aspx

CloverETL Data Profiler
http://www.cloveretl.com/products/profiler

Data Cleaner
http://datacleaner.org/

Datiris
http://www.datiris.com/index.shtml

PitneyBowns Enterprise Data Governance
http://www.pbsoftware.eu/ger/produkte/datenmanagement/datenqualitaet/data-profiling/

ClearInformation Quality Management
http://www.clearinformation.org/index.php/ci-implementation/ci-index-cix/data-profiling

Global IDs Data Profiling and Mapping Suite
http://www.globalids.com/products/product-suites/data-quality-and-verification-dqv

PSTech Data Profiling and Cleansing Services
http://www.pstech.rs/en/services/data-profiling-and-cleansing.html

Metanome
https://hpi.de/naumann/projects/data-profiling-and-analytics/metanome-data-profiling.html

Information Integration
Course Project – Extraction

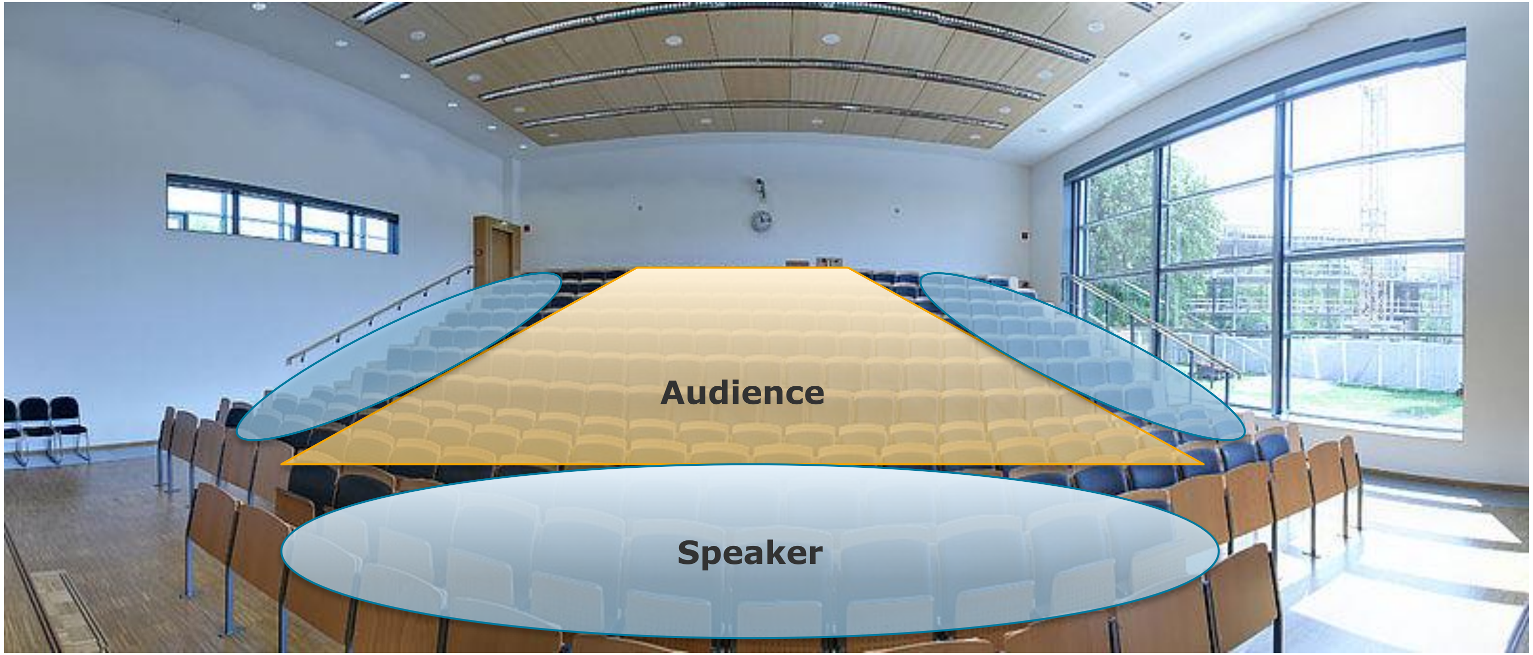
1. Find 2 datasets that each …
   - contain public person data (politicians, athletes, actors, …).
   - contain at least 1 and at most 4 additional entities (party, sport, movie, …).
   - contain more than 4 attributes.
   - originate from different sources / web sites.

2. Design a database schema for each dataset individually that …
   - captures the data as it is (no standardization and no cleansing!).

3. Extract the data you have chosen and insert it into your new schema …
   - using a self written extractor or a tool that you found in the internet.

4. Document your schemata in your presentation slides and …
   - introduce the source datasets (topic, size, source, …).
   - provide the ER-diagram.
   - provide the create table statements (fields, datatypes).
   - provide the add constrain statements (keys, foreign keys).

Course Project – Extraction Presentation



Information Integration

Course Project

Questions to:

Mailing List: <not yet ready>

Thorsten Papenbrock: Email or Office E-2-01.2