

# **TransferMeeting „Big Data und Datenintegration“ Management und Analyse großer Datenmengen**

04. Juli 2013

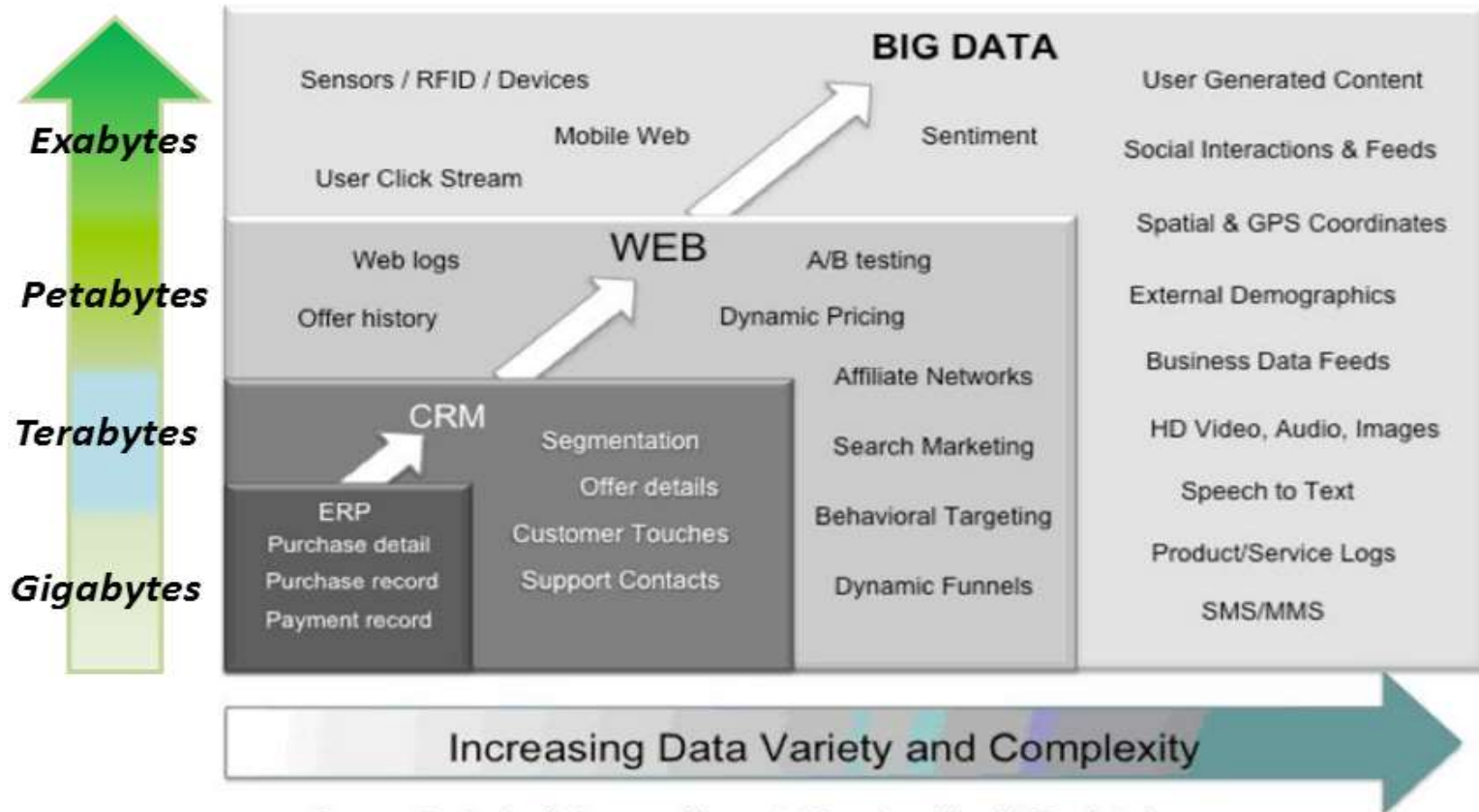
## **Big Data und Datenintegration**

**Prof. Dr. Erhard Rahm**

**<http://dbs.uni-leipzig.de>**



# Massives Wachstum an Daten



Source: Contents of above graphic created in partnership with Teradata, Inc.

Gartner:

- pro Tag werden 2.5 Exabytes an Daten generiert
- 90% aller Daten weltweit wurden in den 2 letzten Jahren erzeugt.

# Datenproduzenten: Soziale Netze, Smartphones, Sensoren ...

UNIVERSITÄT LEIPZIG

**12+ TBs**  
of tweet data  
every day



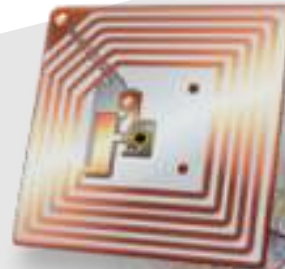
? TBs of  
data every day



**25+ TBs** of  
log data  
every day



**30 billion** RFID  
tags today  
(1.3B in 2005)



**4.6 billion**  
camera  
phones  
world wide



**100s of millions**  
of GPS  
enabled  
devices sold  
annually



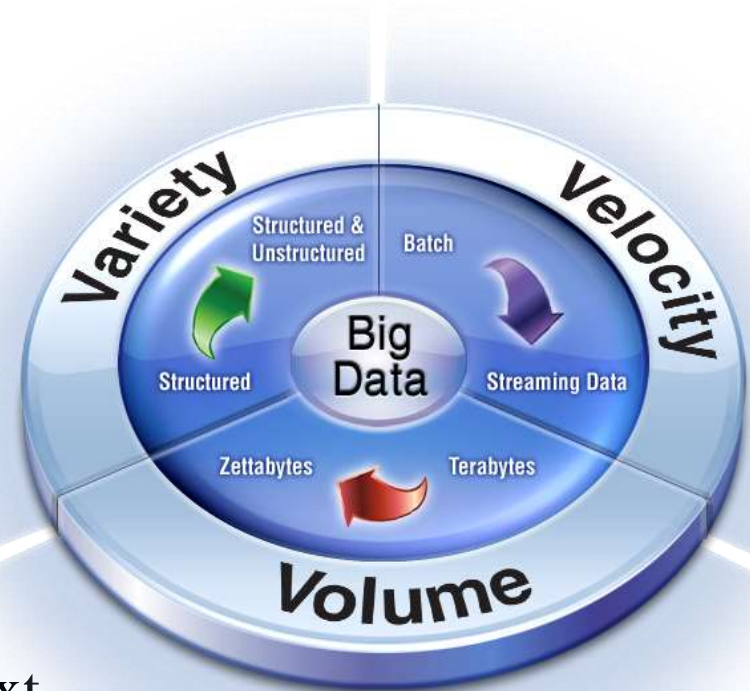
**76 million** smart meters  
in 2009...  
200M by 2014

**2+ billion**  
people on  
the Web by  
end 2011



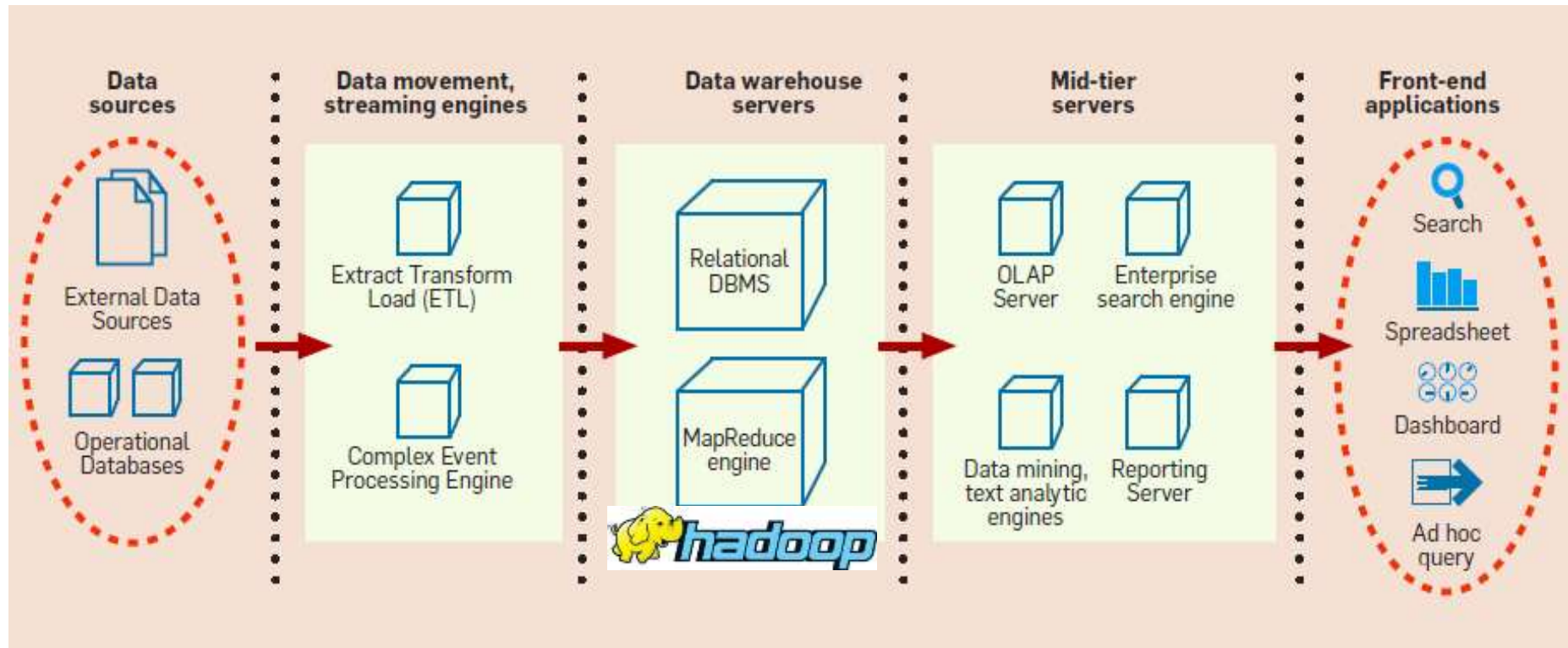
# Big Data Challenges

- Volume** Skalierbarkeit von Terabytes nach Petabytes (1K TBs) bis Zettabytes (1 Milliarde TBs)
- Velocity:** Near-Realtime, Streaming
- Variety** variierende Komplexität: strukturiert, teilstrukturiert, Text
- Veracity:** Vertrauenswürdigkeit
- Value** Erzielen des (wirtschaftl.) Nutzens durch Analysen





# Analyse-Pipeline



- Datenvorverarbeitung und Datenintegration
- Nutzung von hoch skalierbaren Cloud-Infrastrukturen (Hadoop)

# Anwendungsdomänen für Big Data Analytics

Smarter Healthcare



Multi-channel sales



Finance



Log Analysis



Homeland Security



Traffic Control



Telecom



Search Quality



Manufacturing



Trading Analytics



Fraud and Risk



Retail: Churn, NBO



# Forschungsarbeiten

## Web Data Integration Lab (WDI-Lab)

- Semantische Datenintegration von Unternehmensdaten und Web-Daten
- Verfahren zum Abgleich/Matching von Metadaten (Schemas, Ontologien) und Instanzdaten
- Verlinkung von Datenquellen (Linked data)

## • Cloud Data Management / Big Data

- Skalierbares Daten-Management / Last-Balancierung mit Hadoop
- Machine Learning auf Hadoop
- Dedoop: Deduplication based on Hadoop

## • Business Analytics mit NoSQL/Graph-Daten

# Spinoffs



webdata solutions

brush up your data



datavirtuality





- Die Data Virtuality GmbH ist ein Startup-Unternehmen.
- Gegründet in 2012 als Spin-off der Universität Leipzig, davor Forschungsprojekt am Lehrstuhl für Datenbanken.



**Baut ein Data Warehouse automatisiert auf**  
mit Daten aus verschiedenen Quellen



**Zeitersparnis:**

Unsere Software ist sofort einsatzbereit

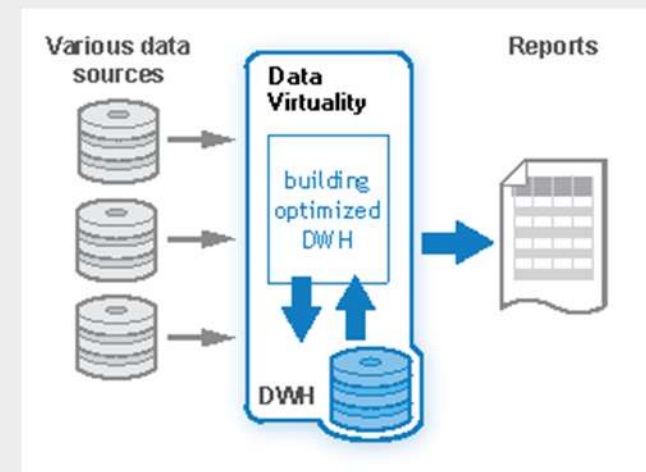


**Kostensparnis:**

automatisierter Aufbau – weniger manuelle Arbeit




**Flexibilität und schnelle Anbindung**  
signifikante Arbeitserleichterung



- Gegründet Jan. 2010
  - Initialförderung durch BMBF (2 Jahre)
  - aktuelle Drittmittelprojekte von EU/DFG/Industrie
- Web-Technologien und Anwendungen
  - Extraktion von (teil-) strukturierten Daten aus Quellen des allg. Web (Suchmaschinen, Portale, Datenbanken) und des „Data Web“ (Linked Data)
  - Datenbereinigung und Matching/Integration heterogener Daten
  - Mashup-Framework WETSUIT (Web EnTity Search and fUslon Tool) zur Realisierung von WDI-Anwendungen

# Integration von Webdaten, z.B. Produktangebote

- Identifikation semantisch äquivalenter Objekte (Objekt-Matching)
- Fusion oder Datenvergleich / Analyse



[Canon \*\*LEGRIA HF S10\*\* Camcorder - 1080p - 8.59 MP - 10 x opt. Zoom](#)  
Flash card, 32 GB SD Memory Card, SDHC-Speicherkarte, HF S10, F/1.8-3.0  
Der HD-Camcorder LEGRIA HF S10 vereint professionelle Leistungsmerkmale mit den Vorzügen von Dual Flash Memory. Moderne Steuerfunktionen ermöglichen die Aufzeichnung in ...  
[Zur Einkaufsliste hinzufügen](#)



[Camcorder Canon \*\*Legria HF S10\*\*](#)  
Canon Legria HF S10 - Camcorder, Video-System: SD-Video, HD-Video, Zoom: 10x optisch, 200x digital, Brennweite: 6,40 mm, 64 mm, Bild-Sensor 1/2,60", ...  
[Zur Einkaufsliste hinzufügen](#)



[Canon \*\*VIXIA HF S10\*\* Camcorder](#)  
Canon VIXIA HF S10 Camcorder SpeicherKarte, Full-HD, NTSC, 10x Optischer Zoom, 0,4 kg  
Der HD-Camcorder LEGRIA HF S10 vereint professionelle Leistungsmerkmale ....  
[Zur Einkaufsliste hinzufügen](#)



[Camcorder Canon \*\*Legria HF S100\*\*](#)  
Canon Legria HF S100 - Camcorder, Video-System: SD-Video, HD-Video, Zoom: 10x optisch, 200x digital, Brennweite: 6,40 mm, 64 mm, Bild-Sensor 1/2,60", ...  
[Zur Einkaufsliste hinzufügen](#)



[Tele Konverter \(Zoom\) 2,0 \*\*CANON LEGRIA HF-S10 HF-S100\*\*](#)  
Tele Konverter (Zoom) 2,0 CANON LEGRIA HF-S10 HF-S100.  
[Zur Einkaufsliste hinzufügen](#)

€955 neu  
von 5 Händlern

[Preise vergleichen](#)

€1.699,00 neu  
Kostenloser Versand  
[Multimedia-Tiefpreise](#)

€823,00 neu  
€829,90 mit Versand  
[fredle-shop](#)

€1.499,00 neu  
Kostenloser Versand  
[Multimedia-Tiefpreise](#)

€58,90 neu  
[Afterbuy-Shops](#)  
[2 Händlerbewertungen](#)

Herausforderungen:

- Schlechte Datenqualität
- Heterogene Repräsentationen
- Fehlerhafte Angaben
- Große Datenmengen
- Verarbeitung in Echtzeit

# Automatische Erkennung von Plagiaten im Internethandel

- Stark zunehmende Verbreitung gefälschter Produkte in Online-Shops und Auktionsplattformen
  - v.a. Luxusgüter wie Parfüm, Handtaschen, Kleidung
- Finanzielle Verluste, Imageverlust, Sicherheitsrisiken



## Desigual MARYLION Women femmes Sac à Main bag

Artikelzustand: **Neu ohne Etikett**

Restzeit: 12T 19Std (01. Mai. 2012 10:09:55 MESZ)

Stückzahl: **1** **4 verfügbar**

**EUR 19,99**

**Sofort-Kaufen**

Auf die Beobachtungsliste

Versand: EUR 7,00 - Sonstiger Versand (Siehe Artikelbeschreibung) |

[Alle Details anzeigen](#)

Artikelstandort: HK, Hong Kong

Versand nach: Weltweit ausgeschlossene Versandorte

Lieferung: Bei Artikeln, die aus dem Ausland verschickt werden, kann der voraus  
Liefertermin nicht genau angegeben werden.

### Business Seller Information

Hai Chuang Ltd

Hedong Qu

Fumin Road Binhetingyuan 32-2-201

300182 Tianjin

V.R China

FAX: 0086(22)87583758

Email: [Liuxiaohui122@gmail.com](mailto:Liuxiaohui122@gmail.com)

### Bewertungsprofil

- kein Original, gefälschte Ware, Kontakt o.k. teilweise Geld erstattet bekommen
- Nicht die Größe, schlecht verarbeitet, es riecht nach modrigem Keller!
- Es ist eine billige Fälschung, keine Originaldesigualbluse, Finger weg

➤ *Hoher Bedarf an semi-automatisierter Identifikation von Angeboten für gefälschte Produkte*

# Vorgehen zur Erkennung von Plagiaten

*Eingabe:* Liste von Produkten eines Herstellers

## Suche

### **Automatisierte Anfragen an Web-Datenquellen:**

- Unscharfe Anfragen an Verkaufsplattformen wie eBay um größtmögliche Abdeckung zu erzielen
- Referenzprodukte von „sicheren Händlern“ wie Otto.de, Amazon

**Datenextraktion:** Verarbeitung der HTML-Seiten und Extraktion der relevanten Daten

**Objekt-Matching:** Abgleich der eingegeben Produkte mit den Ergebnissen

## Clustering

- Einteilung der Angebote in Gruppen von vergleichbaren Produkten
- Ermöglicht Analysen der Preisabweichung zu Angeboten für gleiche Produkte

## Scoring

### **Bewertung der Angebote anhand verschiedener Indikatoren**

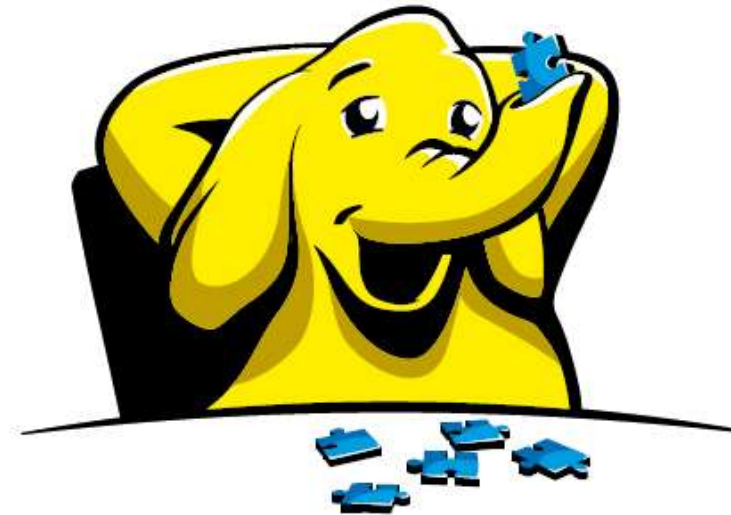
ungewöhnlicher Preis  
Artikelzustand , Ausfälligkeiten in der Beschreibung  
Menge der angebotenen Artikel  
Nutzerkommentare  
Herkunftsland, Zahlungsmethode ...

*Ergebnis:* Liste von verdächtigen Angeboten

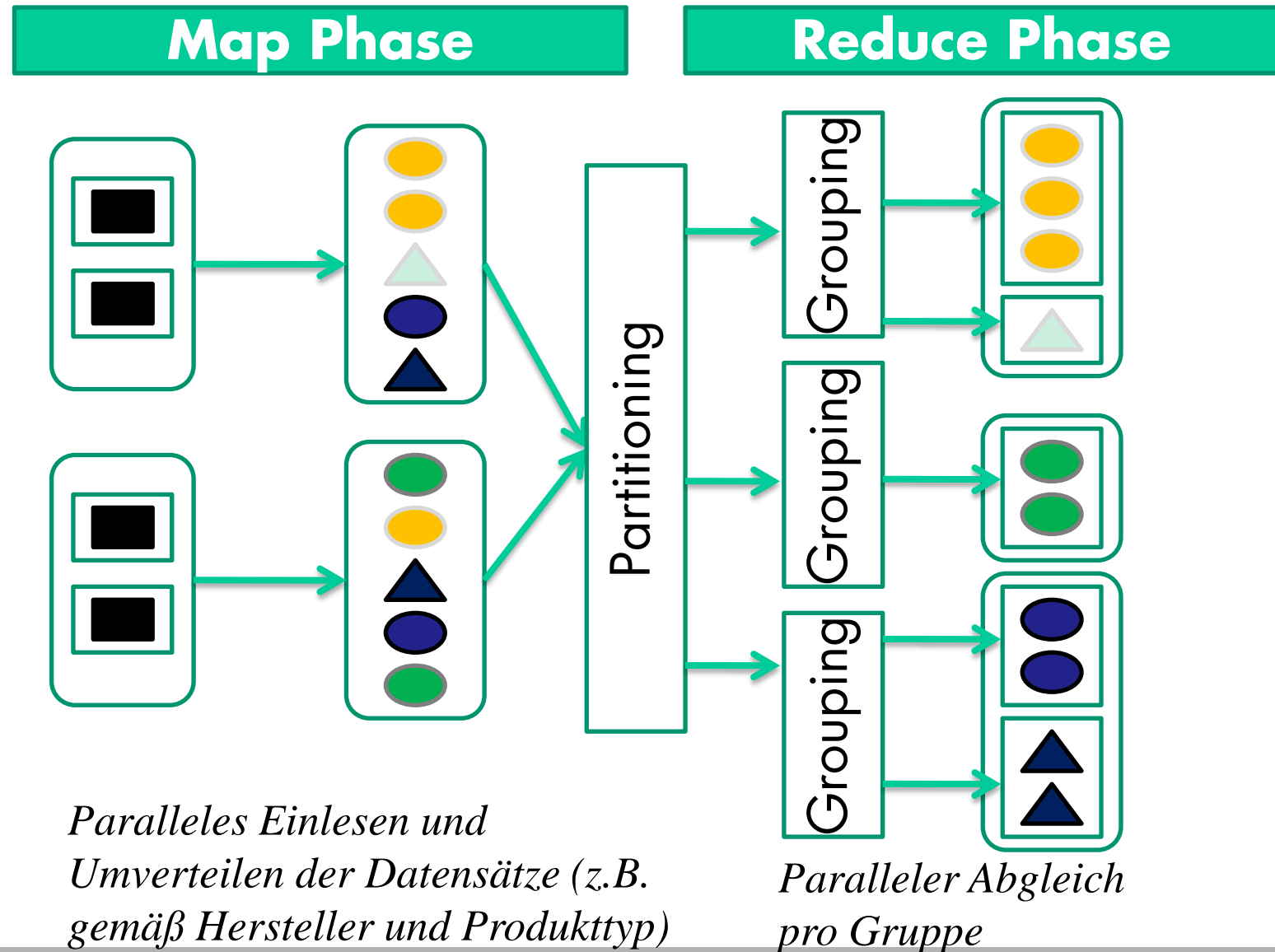


# Dedoop: Efficient Deduplication with Hadoop

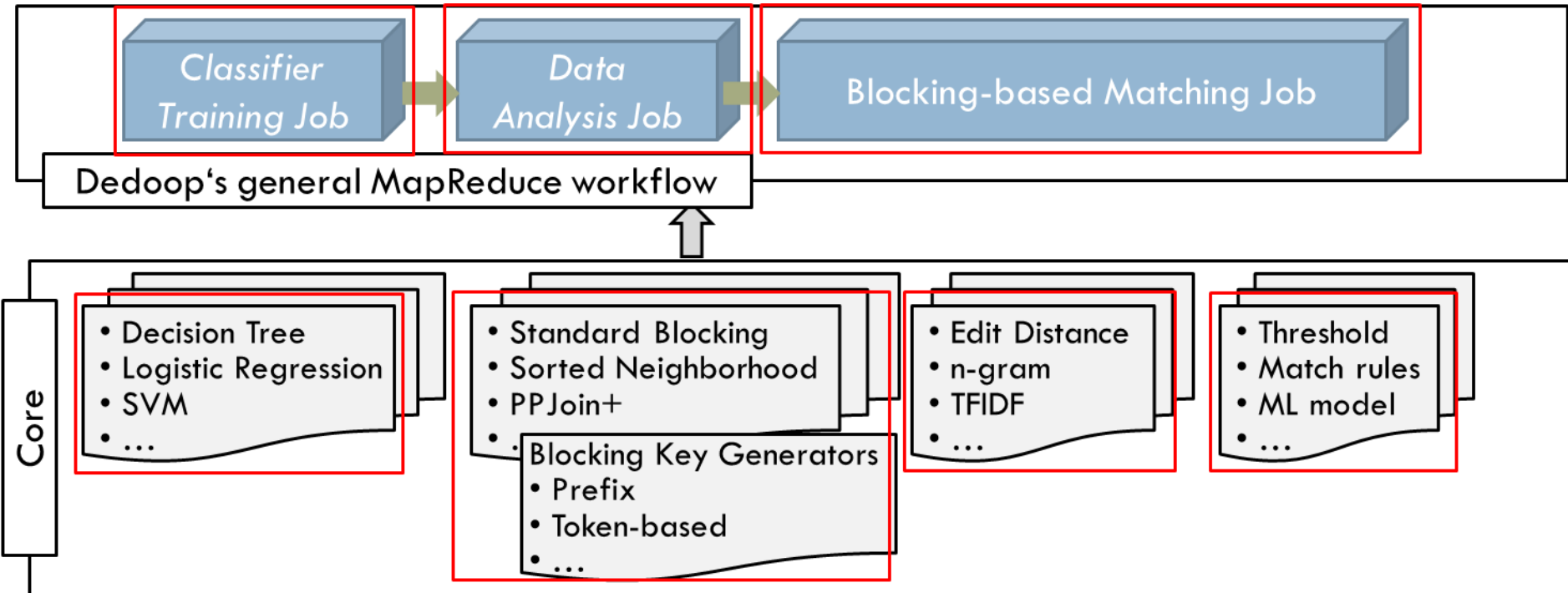
- Parallele Ausführung von Datenintegrations/Match-Workflows mit Hadoop
- Browser-basiertes GUI
- Mächtige Funktionsbibliothek mit
  - vielen Match-Techniken
  - Lernbasierte Konfiguration
- Automatische Generieren und Starten von Map/Reduce-Jobs auf unterschiedlichen Clustern
- Automatische Lastbalancierung
- Monitoring der Ausführung



# Matching mit MapReduce



# Dedoop Überblick



# Browser-basierte Spezifikation

**Dedoop - Efficient Deduplication with MapReduce**

Experiment 1 + Expert Mode

**Hadoop Cluster**

**Running Cluster** Launch EC2 Cluster

Namenode:

Jobtracker:

WebUI port:

Disconnect

**Hadoop Distributed File System**

Name	Size
input_data	
praktikum	
DBLP.txt	362.37K
GoogleScholar.txt	8.83MB
quality_perfect.csv	238.41K
train_500_1.txt	15.01KE
map_reduce	
output	
test	

**Workflow Definition**

Input Data

Mode: ☐ Self-Join ☒ R-S Join

Domain Source:

Range Source:

Id Attribute:

Id Attribute:

dblp\_title:

dblp\_authors:

Attribute Mapping: Attribute 3:

Attribute 4:

Attribute 5:

☒ Normalize attribute values

Output Directory:

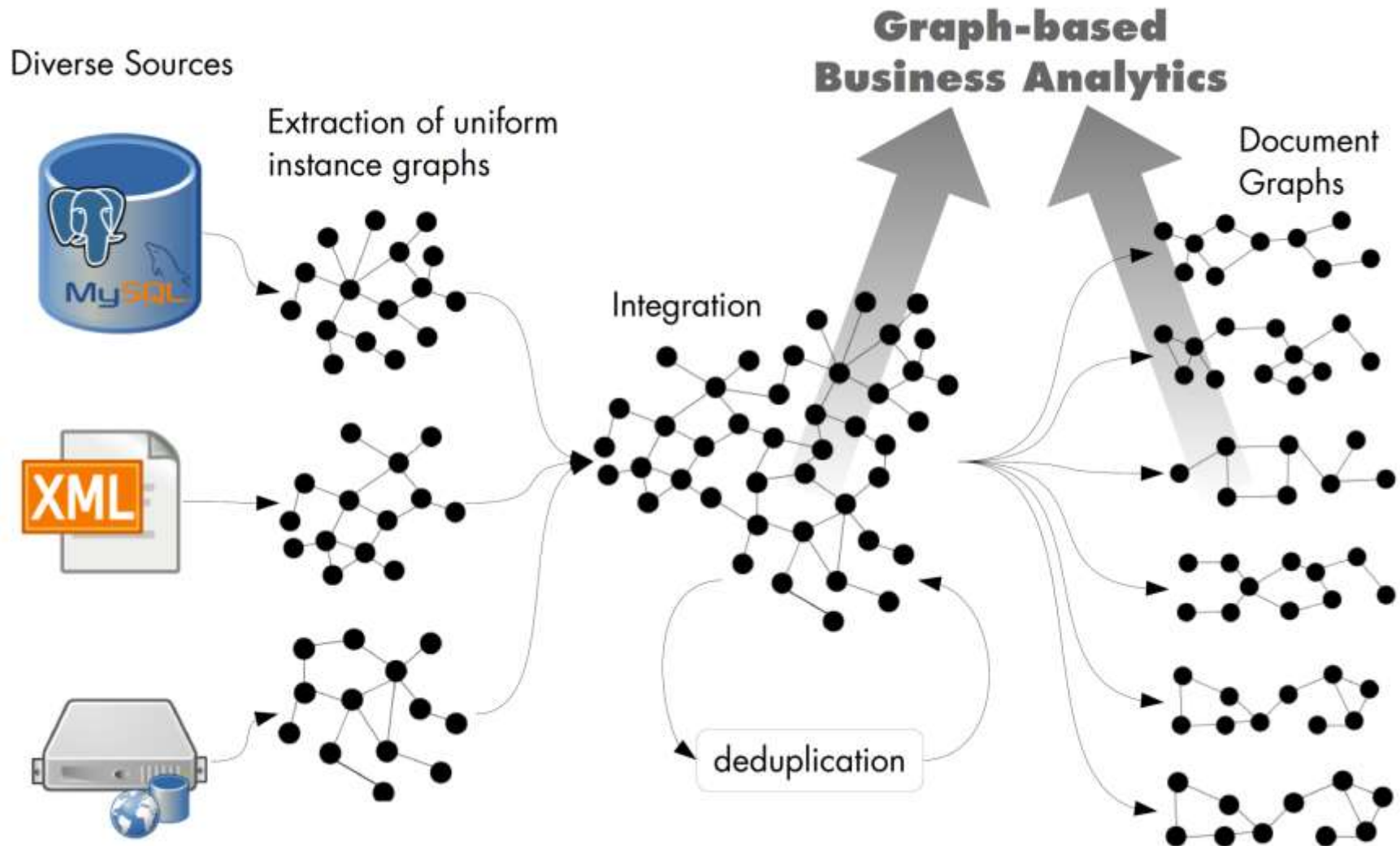
**Data Source definition & File Viewer**

Data Source	Size
hdfs://gkpc3.informatik.uni-leipzig.de/input_data/DBLP.txt	362.37KB
hdfs://gkpc3.informatik.uni-leipzig.de/input_data/GoogleScholar.txt	8.83MB

gs_id	gs_title	gs_authors	Attribute 3	Attribute 4	Attribute 5
0HMk-YUh4i8J	Too Much Middlewar	M Stonebraker	SIGMOD Record,	2002	25
rgzK3sG-rnQJ	A Correctness Proof	M Castro, B Liskov		1999	26
r3sCE4vukG0J	On a stochastic optim	EKP Chong, PJ Ram	Proc. 28th Allerton C		27
7B7KCNJu4j8J	Flight to Objectivity: I	S Bordo			28
wGTOR7lmlYI	Capturing Design Re	M Klein			29

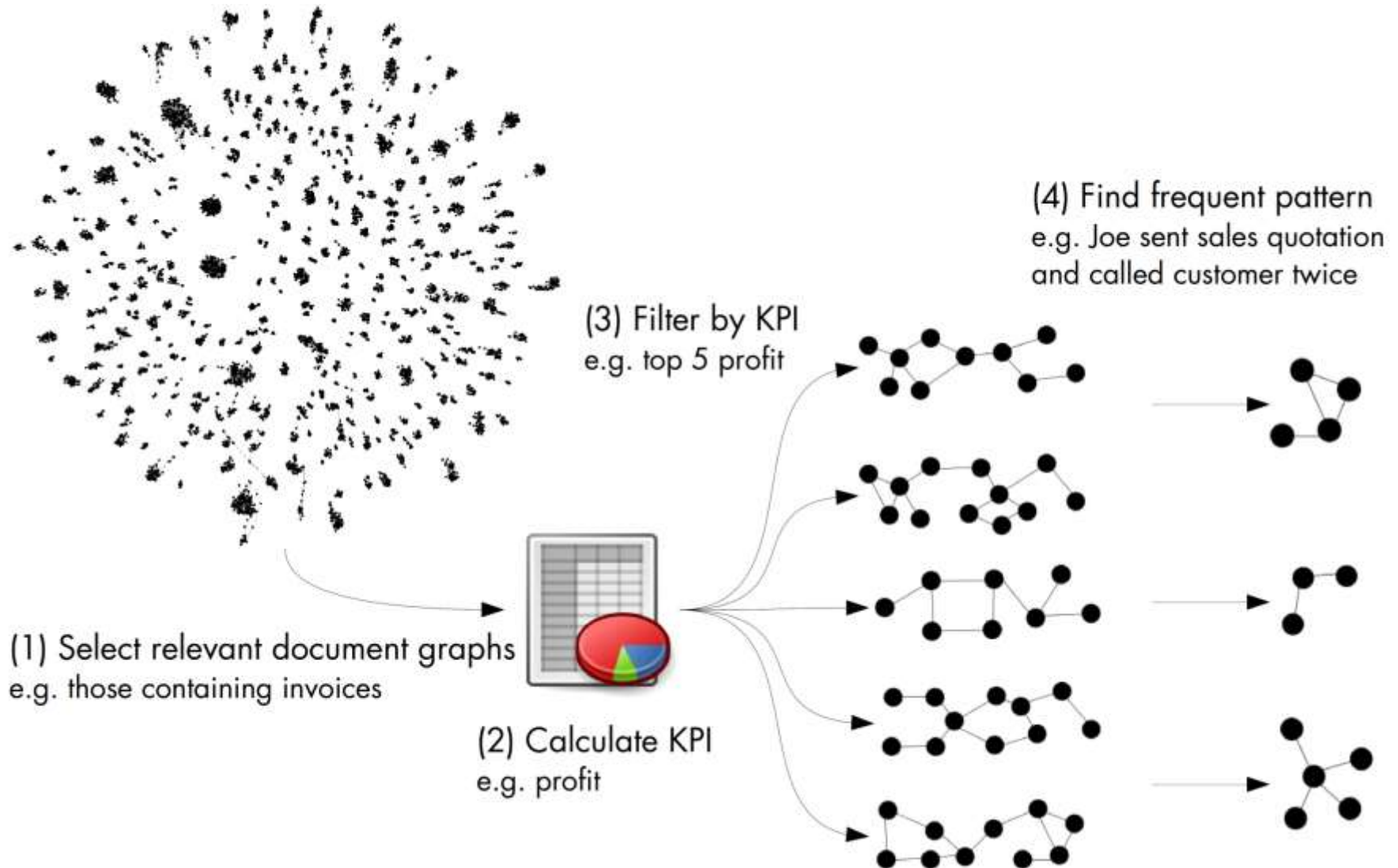
# Graph-basierte Analysen

Framework **BIIG**: Business Intelligence with Integrated Instance Graphs





# BIIG-Analysen



# Unser Angebot

- Beratung / Kooperationen in Bezug auf
  - Big Data und / oder
  - Datenintegration
- Studentische Abschlussarbeiten (Bachelor/Master)
- Gemeinsame Projekte
  - gefördert durch Land/Bund/EU
  - unternehmensfinanziert

# Weitere Informationen

- Persönlich hier auf dem Transfer-Meeting
- Poster/Demos
  - WDI-Lab / Online-Plagiate
  - DEDOOP - Deduplication with Hadoop
  - BIIG
  - Data Virtuality
  - Webdata Solutions
- Web: <http://dbs.uni-leipzig.de>
- Email: [rahm@informatik.uni-leipzig.de](mailto:rahm@informatik.uni-leipzig.de)

**Danke für die Aufmerksamkeit!**

