

Data-Warehouse-Praktikum

WS 18/19

Universität Leipzig, Institut für Informatik

Abteilung Datenbanken

Prof. Dr. E. Rahm

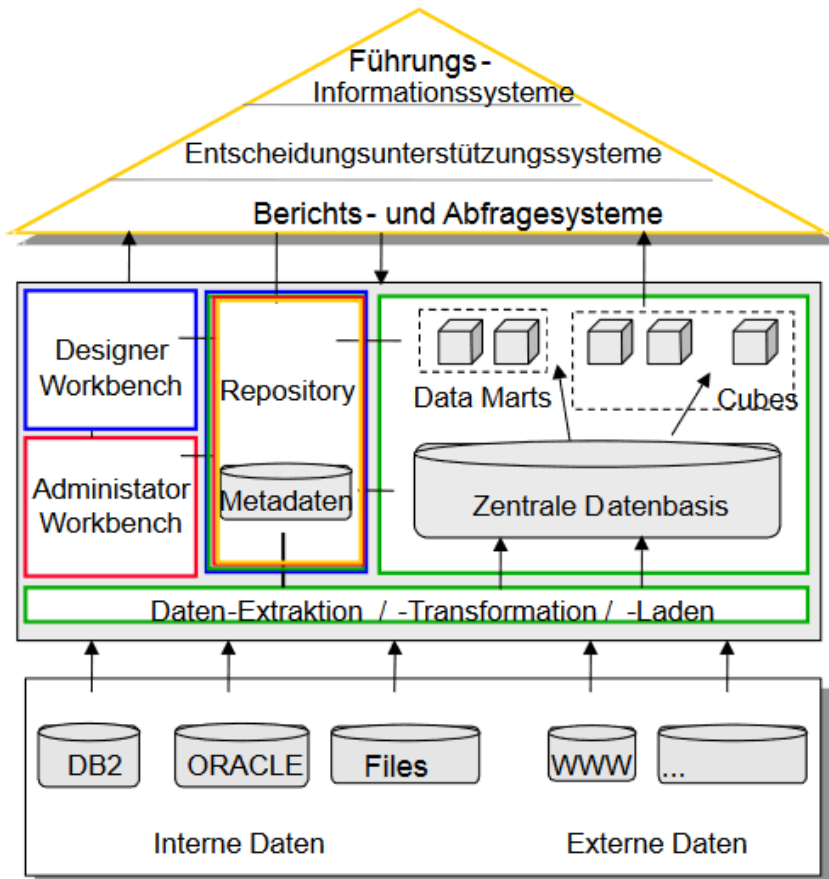
V. Christen, M. Franke, Z. Sehili

{christen, franke, sehili}@informatik.uni-leipzig.de

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

Data-Warehouse

- Ausgangsproblem
 - Viele Unternehmen haben Unmengen an Daten, ohne daraus ausreichend Informationen und Wissen für kritische Entscheidungsaufgaben ableiten zu können

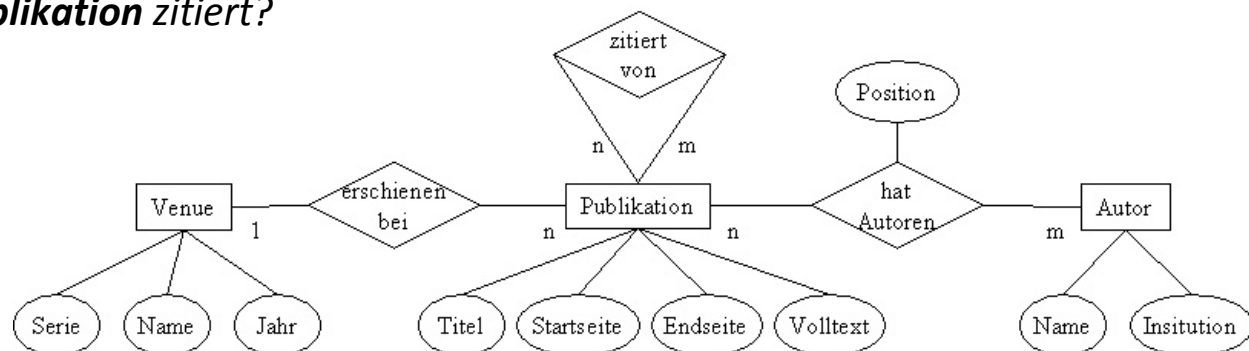


ETL-Prozess:

- **Extraktion**
 - Laden der Quelldaten in temporären Arbeitsbereich
- **Transformation**
 - Anpassung an das Zielschema
 - Datenbereinigung und Integration
- **Laden**
 - Data Cube Erstellung

Szenario: Zitationsanalyse

- In wissenschaftlichen Arbeiten werden andere Arbeiten zitiert
- **Anzahl** der Zitierungen als Indikator für den wissenschaftlichen Einfluss (Impact) und Qualität
 - *Wie häufig wird eine **Publikation** zitiert?*
 - *Wie häufig werden Publikationen des **Venues** (Konferenz oder Journal) im Durchschnitt zitiert?*
 - *Wie ist die durchschnittliche Zitierungszahl von **Autoren**?*
- Beziehungen zwischen Personen, Institutionen, Publikationen und Fachbereichen
 - *Welche **Autoren** zitieren welche anderen **Autoren**?*
- Verlagerung von Forschungsschwerpunkten, Trend-Themen
 - *Wann wird ein **Publikation** zitiert?*



Datenquellen

- **DBLP Bibliography:**
 - Manuelle gepflegte Website, die komplette Listen verschiedener Venues aus dem Informatik-Bereich enthält.
- **ACM Digital Library:**
 - Portal der Association for Computing Machinery
 - enthält ebenfalls komplette Listen verschiedener Venues
- **Google Scholar:**
 - Suchmaschine für wissenschaftliche Publikationen
- Relevante Teilmenge der Daten steht als CSV- und XML-Dateien zur Verfügung

Aufgaben: Inhaltlich

1. Datenimport

- Import der XML- und CSV-Dateien
- Datenextraktion mittels TSQL
- Relationale Speicherung der Daten dem **Zielschema entsprechend** (Zusatzinformationen sollen beibehalten)

2. Data Cleaning

- Objekt-Matching: Erkennen gleicher Publikationen in verschiedenen (oder gleichen) Datenquellen
- Daten-Normalisierung: Normalisierung der Institutionsnamen
- Ableitung neuer Daten: Identifikation von Selbstzitationen

3. Cube-Erstellung, OLAP und Data Mining

- Star-Schema-Erstellung und Datenimport
- OLAP-Analyse, MDX-Anfragen
- Data Mining: Assoziationsregeln zur Bestimmung *ähnlicher Venues*

Aufgaben: Organisatorisch

- Realisierung mittels *Microsoft SQL Server Business Intelligence Development Studio*
 - Drag&Drop-Workflow-Erstellung (keine direkte Programmierung)
 - Per Remote-Desktop-Verbindung: **wdiserv3.informatik.uni-leipzig.de**
 - Client-Anwendung für zentralen Datenbankserver: SQL Server 2008 auf **windorf.informatik.uni-leipzig.de**

*Für Experimentierfreudige oder Open-Source-Verfechter: Nutzung von **Pentaho** (Mondrian / Kettle)*
<https://community.hitachivantara.com/docs/DOC-1009855-data-integration-kettle>

- Jeder Aufgabe ist ein Tutorial zugeordnet
 - Beschreibung der Aufgabe
 - Grundlegende Vorgehensweise (inkl. Screenshots) & Hinweise
- Software-Ergebnis sind ausführbare Projekte, welche im **Testat** ausgeführt/begutachtet werden
 - Terminabsprache rechtzeitig individuell mit Betreuer per E-Mail
 - Deadlines siehe Webseite (Testat 1 bis zum 23. November)

Organisatorisches

- **Ziel:** Realisierung eines *typischen* DWH-Projekts
 - Kennenlernen der *echten, praktischen* DWH-Probleme
- **Zielgruppe:**
 - Master-Studierende der Informatik, welche im SS18 VL Data Warehousing besucht haben
 - Interessierte (nachrangig)
- **Kenntnisse :**
 - Notwendig: VL *Data Warehousing*
 - Hilfreich: VL Datenintegration, VL Data Mining, DB-Praktikum
 - Skripte zum Nacharbeiten im Netz
- **Ablauf:**
 - Gruppenarbeit mit 2 Studierenden pro Gruppe (max. 9 Gruppen)
 - Bearbeitung von 3 Aufgaben → jeweils Testat
- **Aufgabenstellung und Informationen:**
 - <https://dbs.uni-leipzig.de/study/2018ws/dwhprak>