

# Data-Warehouse-Praktikum

Einführungsveranstaltung WS 17/18

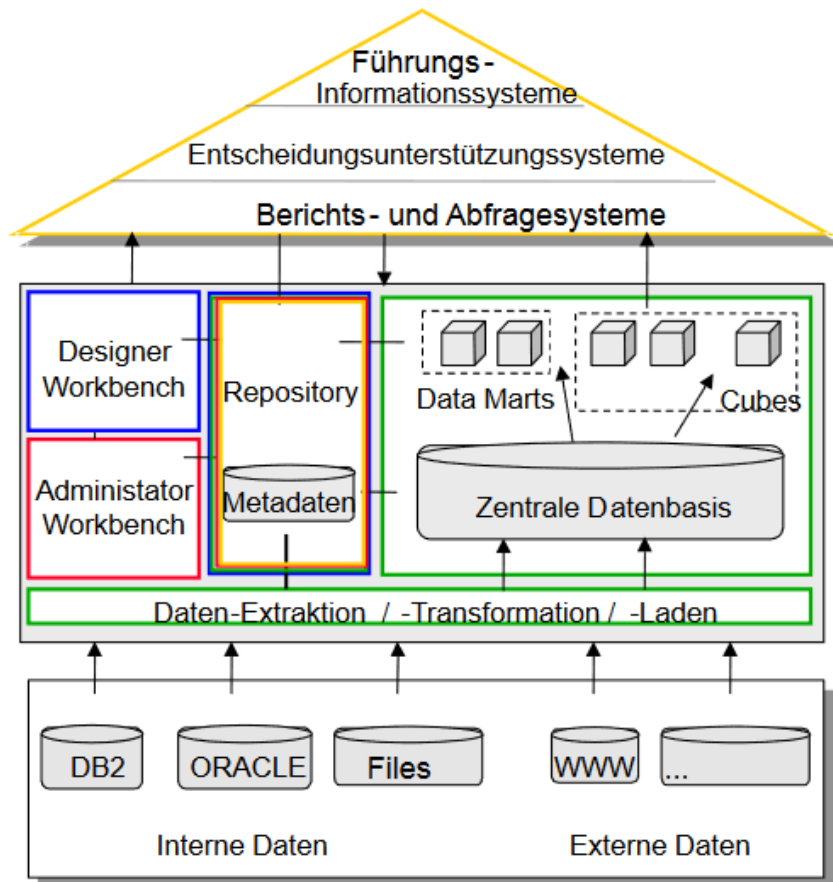
23.10.2017

Victor Christen, Martin Franke, Ziad Sehili  
Abteilung Datenbanken, Universität Leipzig  
<http://dbs.uni-leipzig.de>

# Datawarehouse

- Ausgangsproblem

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

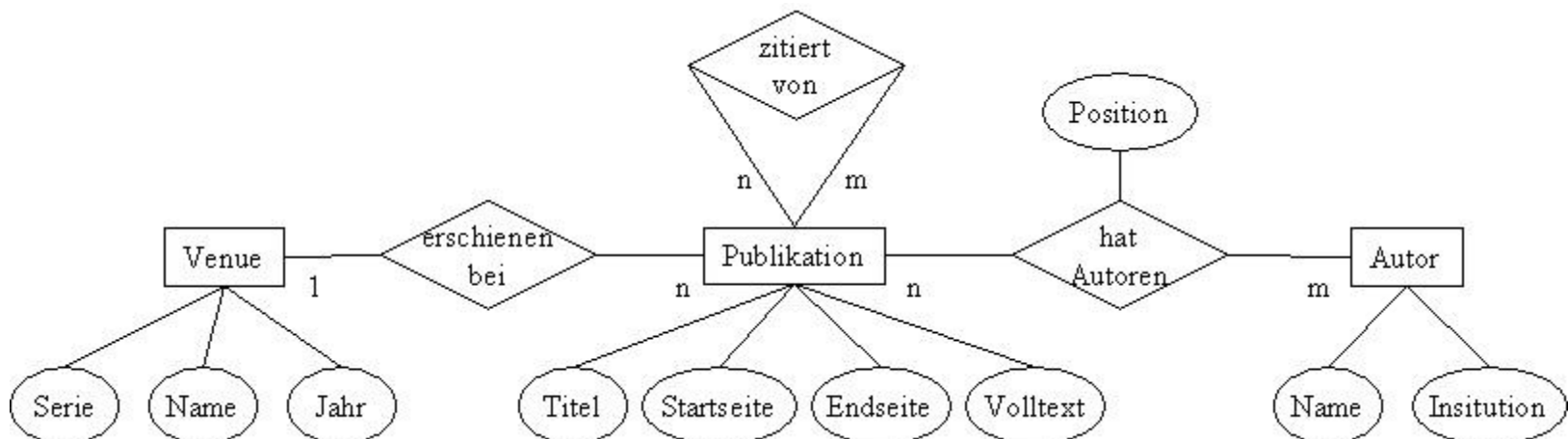


## ETL – Prozess

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

# Szenario: Zitierungsanalyse

- Innerhalb wissenschaftlicher Arbeiten werden andere Arbeiten zitiert
- Anzahl der Zitierungen charakterisiert wissenschaftlichen Einfluss
  - Wie häufig wird Publikation X zitiert?
  - Wie häufig werden Publikationen des Venues (= Konferenz oder Journal) Y im Durchschnitt zitiert?
  - Wie ist die durchschnittliche Zitierungszahl von Autor Z?



# 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

## 2. Data Cleaning

- Objektkonsolidierung: Erkennen gleicher Publikationen in verschiedenen (oder gleichen) Datenquellen
- Datennormalisierung: Normalisierung der Institutionsnamen
- Ableitung neuer Daten: Identifikation von Selbstzitierungen

## 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 "SQL Server Business Intelligence Development Studio"
  - "Drag&Drop"-Workflow-Erstellung (keine Programmierung)
  - Per Remote Desktop Verbindung:  
**wdiserv3.informatik.uni-leipzig.de**
  - Client-Anwendung für zentralen Datenbankserver SQL Server 2008 auf **windorf**
- Jeder Aufgabe ist ein Tutorial zugeordnet
  - Beschreibung der Aufgabe
  - Grundlegende Vorgehensweise (inkl. Screenshots)
  - Weitere Hinweise
- Software-Ergebnis sind ausführbare Projekte, welche im Testat ausgeführt/begutachtet werden
  - Terminabsprache per E-Mail
  - Deadlines siehe Webseite

# Organisatorisches

- Ziel: Realisierung eines "typischen" DWH-Projekts
  - Kennenlernen der "echten, praktischen" DWH-Probleme
- Zielgruppe
  - Informatik-Studenten (Master, Diplom)
  - Interessierte
- Kenntnisse
  - Vorlesung "Data Warehousing" oder äquivalente Vorkenntnisse nötig
  - Vorlesung "Datenintegration" hilfreich
  - Skripte zum Selbststudium/Nacharbeiten im Netz
- Ablauf
  - Gruppenarbeit mit 2/3 Studenten pro Gruppe
  - Bearbeitung von 3 Aufgaben → jeweils Testat → Schein
- Aufgabenstellung und Informationen
  - <https://dbs.uni-leipzig.de/study/2017ws/dwhprak>

