Data-Warehouse-Praktikum

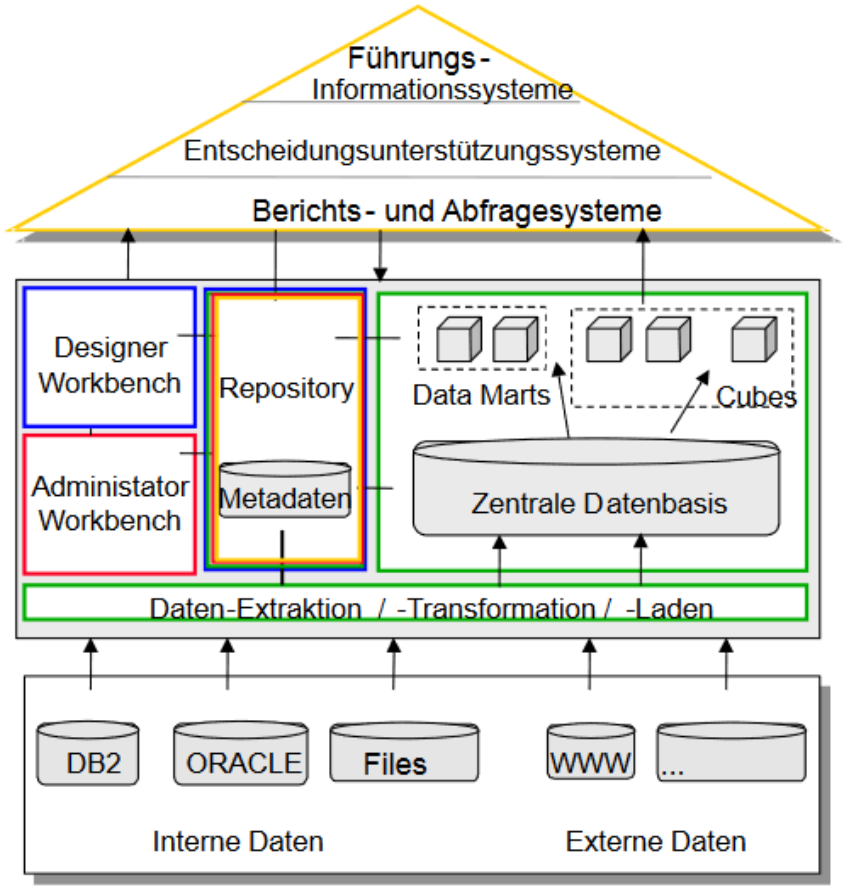
WS 19/20

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

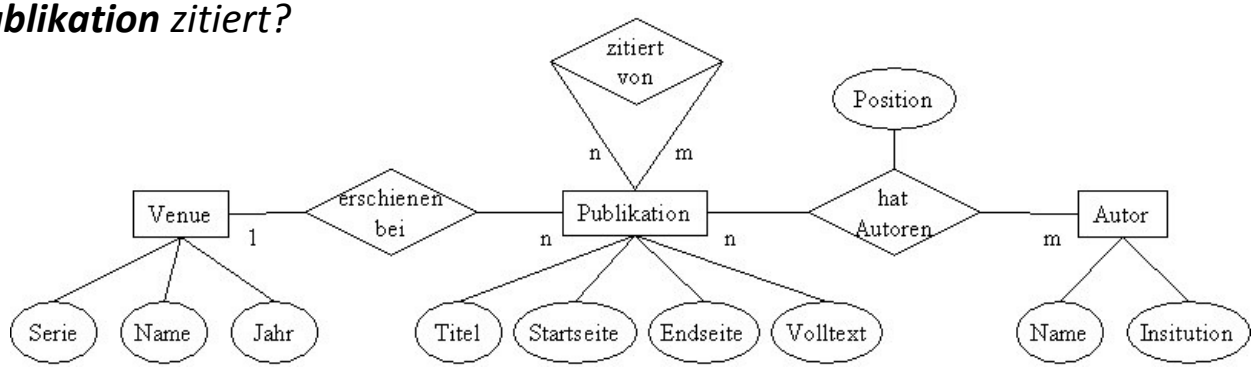


ETL-Prozess:

- **E**xtraktion
  - Laden der Quelldaten in temporären Arbeitsbereich
- **T**ransformation
  - Anpassung an das Zielschema
  - Datenbereinigung und Integration
- **L**aden
  - 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 eine **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 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

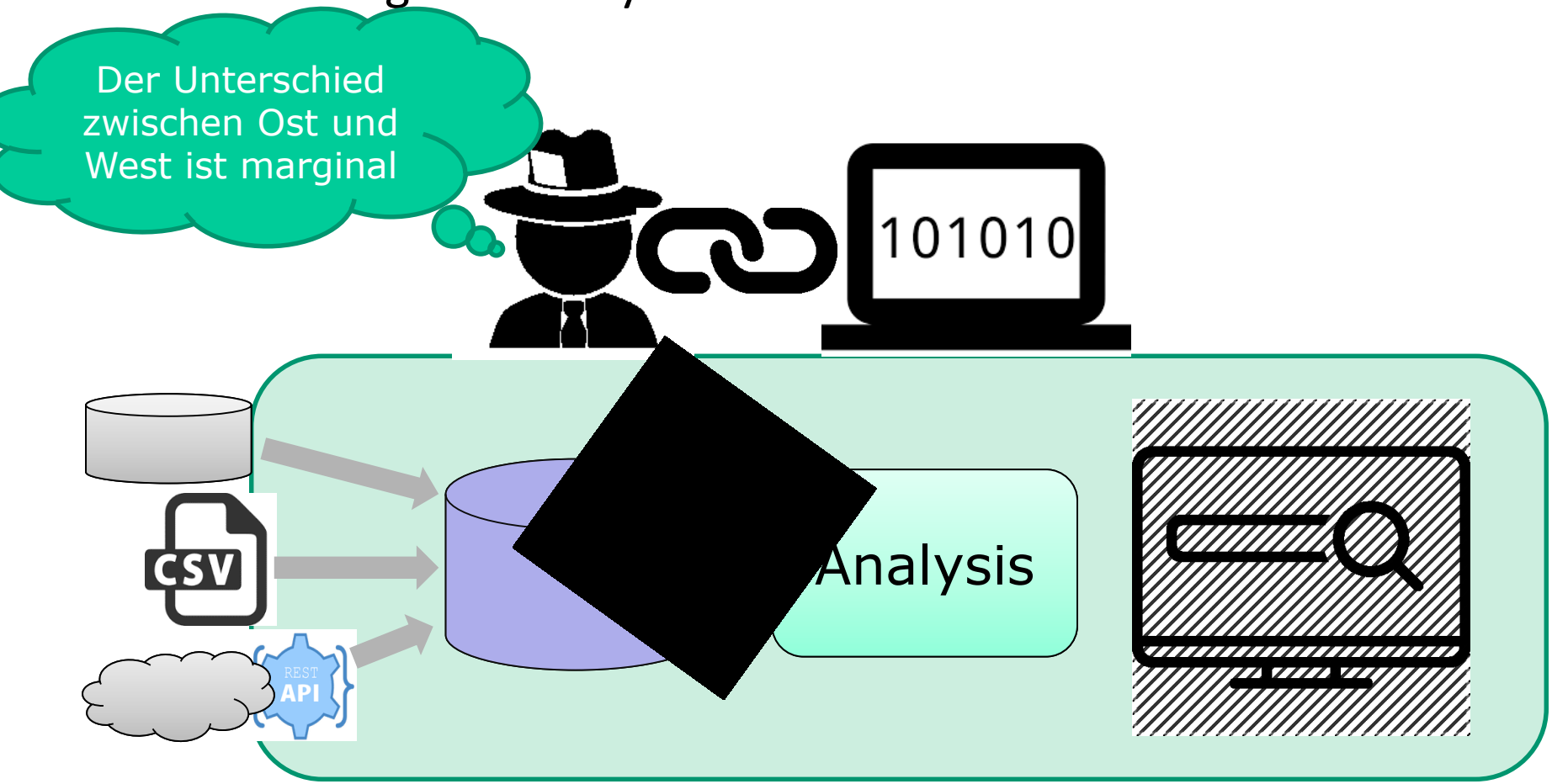
- Realisierung mittels *Microsoft SQL Server Business Intelligence Development Studio*
  - Drag&Drop-Workflow-Erstellung (keine direkte Programmierung)
  - Per Remote-Desktop-Verbindung: **wdiserv3.informatik.uni-leipzig.de**
  - Client-Anwendung für zentralen Datenbankserver: SQL Server 2014 auf **windorf.informatik.uni-leipzig.de**
- 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 29. 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: *Data Warehousing* Verständnis
  - Hilfreich: VL Datenintegration, VL Data Mining, DB-Praktikum
  - Skripte zum Nacharbeiten im Netz
- **Ablauf:**
  - Gruppenarbeit mit 2 Studierenden pro Gruppe
  - Bearbeitung von 3 Aufgaben → jeweils Testat
- **Aufgabenstellung und Informationen:**
  - https://dbs.uni-leipzig.de/study/2019ws/dwhprak

Data-Mining in der Praxis

- Datenjournalismus ist die Aufbereitung, Analyse und Publikation öffentlich zugänglicher Daten
- Aufbereitung und Analyse erfordert automatisierte Verfahren



Inhalt

- Kooperation mit Masterstudenten des Studiengangs Datenjournalismus bei dem MDR-Projekt „Deutschland-Doppel“ anlässlich 30 jähriger Wiedervereinigung
- Entwicklung einer Webapplikation für die historische Analyse deutscher Städte der BRD und DDR
  - Data-Warehouse
    - ETL- Prozess
  - Data-Mining
    - Ähnliche Städte → Clustering, TopK-Selektion, Outlier- Detection
  - Visualisierung
    - Interaktive Webanwendung, wie z.B. Highlighting ähnlicher Städte bei Selektion einer Stadt, Time-Slider mit Anzeige ähnlicher Städte

Ablauf

Termine des Datenjournalismus Seminars

- Zeppelinhaus 3.18, Nikolaistraße 25, mittwochs, 11:15-12:45 Uhr

Datum	Thema
23.10.2019	Konkretisierung des Projektauftrags
13.11.2019	Finalisierung des Projektmanagement-Plans
27.11.2019-15.01.2020	Projektarbeit

- Absolvierung und Präsentation der 3 Testat bei der Abteilung Datenbanken

Testate

Testat 1

- Konzeptualisierung eines standardisierten Datenmodells
- Extraktion, Transformation, Speicherung der Daten dem Datenmodell entsprechend

Testat 2

- Implementierung und Ausführung von Clustering, Top-K Selektion Verfahren. Der Umfang der Verfahren kann sich ändern.

Testat 3

- Realisierung einer interaktiven Webapplikation für die Präsentation der Analysen