UNIVERSITÄT LEIPZIG

DATA-WAREHOUSE-PRAKTIKUM

Universität Leipzig, Institut für Informatik
Abteilung Datenbanken
Prof. Dr. E. Rahm
V. Christen, L. Lange, F. Rohde, B. Uhrich

{christen, lange, rohde, uhrich}@informatik.uni-leipzig.de
http://dbs.uni-leipzig.de

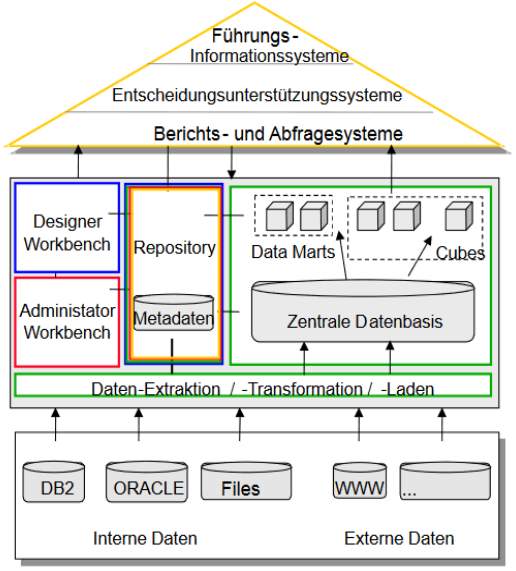
UNIVERSITÄT LEIPZIG

ORGANISATORISCHES

- Ziel: Realisierung eines typischen DWH-Projekts
  - Kennenlernen der echten, praktischen DWH-Probleme
- Zielgruppe:
  - Master-Studierende Data Science
  - Kenntnisse :
    - Notwendig: Data Warehousing Verständnis
    - Hilfreich: VL Data Mining, DB-Praktikum
- Ablauf:
  - Gruppenarbeit mit 2 Studierenden pro Gruppe
  - Bearbeitung von 3 Aufgaben → jeweils Testat
- Aufgabenstellung und Informationen:
  - Moodle

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

UNIVERSITÄT LEIPZIG

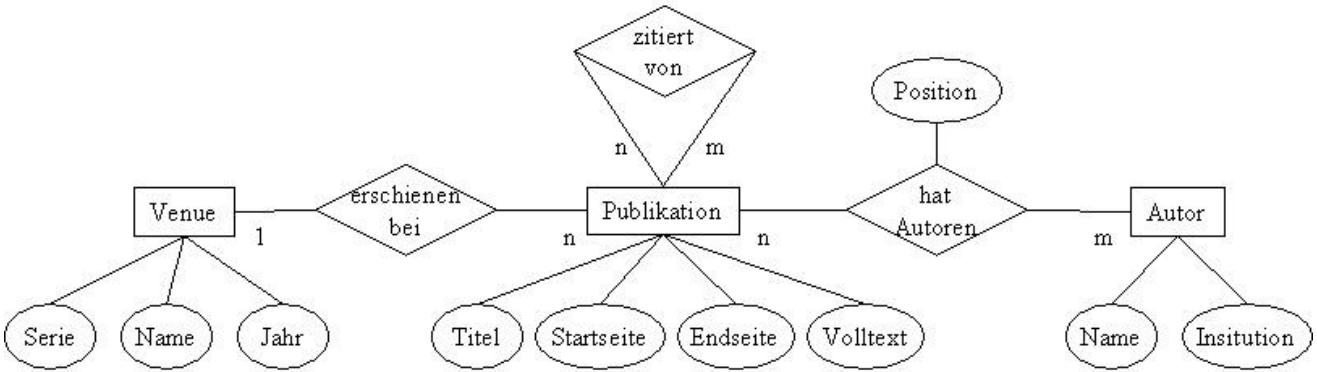
SZENARIO: ZITATIONSANALYSE

- 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

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

ZIELSCHEMA


zitiert
von
n
m
Position
Venue
1
erschienen
bei
n
Publikation
n
hat
Autoren
m
Autor
Serie
Name
Jahr
Titel
Startseite
Endseite
Volltext
Name
Insitution

UNIVERSITÄT LEIPZIG

AUFGABEN

1. Datenimport

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

2. Data Cleaning

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

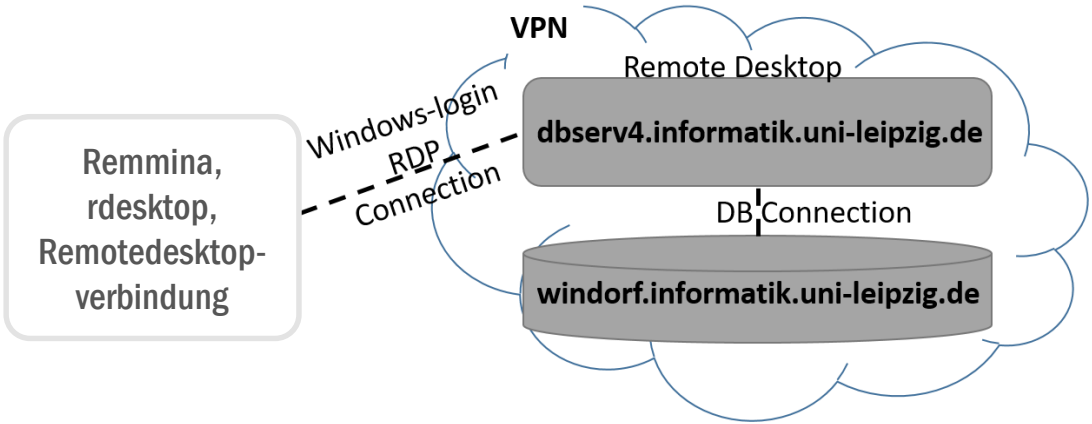
3. Cube-Erstellung, OLAP und Data Mining

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

UNIVERSITÄT LEIPZIG

ENTWICKLUNGSUMGEBUNG

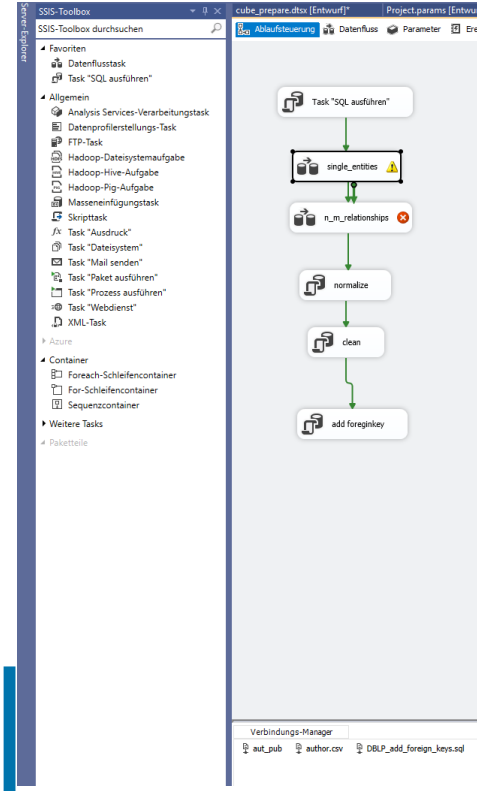
- Bereitstellung vollständiger Entwicklungsumgebung
  - Windows-Accounts nach Themeneinschreibung



- Vorteil: kein Installationsaufwand, bessere Hilfestellung mgl.
- Nachteil: steile Lernkurve, Tool “gewöhnungsbedürftig”

UNIVERSITÄT LEIPZIG

REALISIERUNG

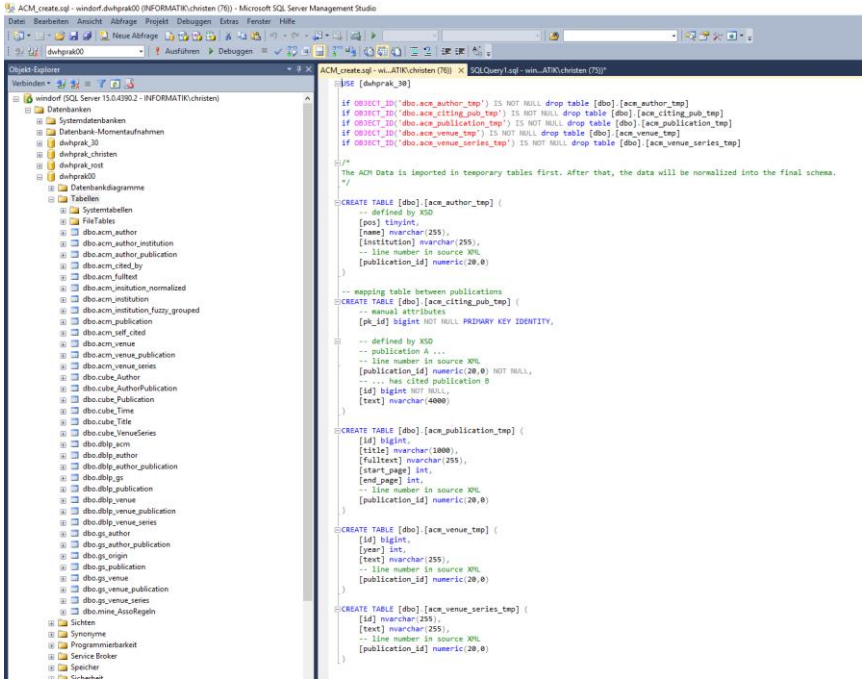


- *SQL Server Data Tools innerhalb von Visual Studio*
  - Drag&Drop-Workflow-Erstellung (keine direkte Programmierung)
  - Per Remote-Desktop-Verbindung: **dbserv4.informatik.uni-leipzig.de**
  - Client-Anwendung für zentralen Datenbankserver: SQL Server 2019 auf **windorf.informatik.uni-leipzig.de**

UNIVERSITÄT LEIPZIG

REALISIERUNG

- *SQL Management Studio*
  - Erstellung von SQL-Skripten (Tabellenerstellung, Datentransformation, etc.)



UNIVERSITÄT LEIPZIG

ORGANISATORISCH

- 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 Moodle(Testat 1 bis zum 13. Dezember)