Prof. Dr.-Ing. Thomas Kudraß
Dipl.-Math. Dörte König
HTWK Leipzig, F IMN

Projekt zur Lehrveranstaltung "Informationssysteme"

Das Projekt ist in drei Teile aufgeteilt, die den Phasen eines Data-Warehouse-Projekts entsprechen. Die Bewertung des Projekts geht zu 50% in die Gesamtnote des Faches „Informationssysteme" ein.

1. Teil: Datenextraktion aus dem Web und Datenimport

- Import der Rohdaten aus dem Web mit Hilfe der Browser-Erweiterung Chickenfoot zum automatisierten Browsen in ausgewählten Webseiten
- Optional: Nutzung von Webservice-Technologien zum Zugriff auf externe Datendienste (z.B. Google, Amazon)
- Umwandlung der Daten in ein XML-Format (z.B. mittels JavaScript, XPath, FireBug PlugIn) und Import in die relationale Datenbank
- Extraktion der für die Analyse benötigten relevanten Merkmale

2. Teil: Data Cleaning / Laden

- Data Cleaning, d.h. Objektkonsolidierung, Datennormalisierung, Ableitung neuer Attribute
- Beispiele für mögliche Bereinigungen: Angleichung unterschiedlicher Schreibweisen von Namen, Erkennung von Duplikaten, Inkonsistenzen
- Ausführung und Diskussion der Data-Cleaning-Schritte hinsichtlich Performanz und Datenqualität
- Laden des ODS bzw. des Data Warehouse

3. Teil: OLAP und Business Intelligence

- Definition und Erstellung eines Data Cubes, inklusive OLAP-Analysen und mehrdimensionaler Abfragen
- Nutzung von Oracle-OLAP-Technologien
- Modellierung von hierarchischen Dimensionen
- Ergebnisse:
    - Aufbau des Data-Warehouse-Schemas (Dimensions- und Faktentabellen mit allen Attributen)
    - Füllen der Dimensions- und Faktentabellen im Analytic Workspace
    - Erstellung der Data Cubes
    - Ausführung interaktiver OLAP-Queries basierend auf Cubes
    - Visualisierung der Anfrageergebnisse (Diagramme)

Der ETL-Prozess kann mit dem Oracle Warehouse Builder modelliert werden.

Zum Projekt ist eine kleine Dokumentation zu erstellen. Diese sollte folgende Bestandteile beinhalten:

Teil 1:

Kurze Beschreibung der gewählten Data-Warehouse-Anwendung (Weltausschnitt). Dazu zählen die wichtigsten Informationsobjekte sowie die geplanten Auswertungen.

Konzeptionelle Modellierung der Datenquellen (als Datenbank oder Dokument)

Aspekte der Datenqualität (Diskussion)

Teil 2:

Konzeptionelles Modell des Operational Data Store (ERD)

Logisches Modell des Operational Data Store (Relationenschema)

Teil 3:

Data-Warehouse-Modell als multidimensionales Datenmodell

Definition von mindestens 3 unterschiedlichen OLAP-Queries

Darüber hinaus relevante Metadaten sind jeweils in den einzelnen Teilen anzugeben.

Teil 1 und Teil 2 zusammen werden in schriftlicher Form als Meilenstein I abgenommen (Termin KW 22). Meilenstein II bildet die Endabnahme am Rechner (Termin KW 26)

Das Projekt ist in Zweiergruppen zu bearbeiten.

Mögliche Anwendungs-Szenarien für den Aufbau eines Data Warehouse:

Fallstudie 1: Leipzig-Event--Data Warehouse

Aufbau eines Data Warehouse über Veranstaltungen und Events in Leipzig, Kapazitäten von Hotels und Gaststätten und anderen Serviceanbietern mit dem Ziel einer besseren Abstimmung und Ressourcenplanung zwischen den Beteiligten.

Öffentlich zugängliche Datenquellen:

- Veranstaltungen (Events):
    - Veranstaltungsdatenbank von Zeitschriften, Magazinen und Portalen (Leipzig live, LVZ, Kreuzer, meinestadt.de, port01)
    - offizielles Portal der Stadt Leipzig
    - Veranstaltungskalender der Stadt Leipzig
- Locations:
    - Locationagent
    - livegigs.de
- Unterkünfte:
    - hrs.de, hotel.de, diverse regionale Portale
- Gastronomie:
    - Webauftritte von Restaurants und Catering-Firmen
- Sonstige Serviceanbieter:
    - Security-, Escort- und Carservice

Informationsbedarf:

- Veranstaltungen (Events):
    - Name, Veranstaltungstyp (z.B. Konferenz, Party, Konzert,…)
    - Veranstaltungszeitraum, Veranstaltungszeit
    - Veranstaltungsort (Adresse oder Location)
    - Kontaktdaten Anbieter
    - Eintrittspreis
- Locations:
    - Name
    - Adresse (einschl. Stadtteil)
    - Kontaktdaten (Telefon, Fax, E-Mail, Web)
    - Erreichbarkeit
- Unterkünfte:
    - Name des Hotels
    - Adresse (einschl. Stadtteil)
    - Kontaktdaten (Telefon, Fax, E-Mail, Web)
    - Bettenkapazität

    - Erreichbarkeit
- Gastronomie:
    - Name der Einrichtung
    - Adresse (einschl. Stadtteil)
    - Kontaktdaten (Telefon, Fax, E-Mail, Web)
    - verfügbare Plätze
    - Angebot (Speisen, Getränke)
- Sonstige Serviceanbieter:
    - Name der Einrichtung
    - Adresse
    - Kontaktdaten (Telefon, Fax, E-Mail, Web)
    - Verfügbarer Service

Mögliche Auswertungen

- Ressourcen- und Kapazitätsvergleich
- langfristige Übersicht über Events (Aufdeckung von Spitzenlasten und Konkurrenz-situationen in der Zukunft)
- Auswertungen in verschiedenen Dimensionen:
    - Veranstaltungstyp
    - zeitbezogen
    - ortsbezogene Auswertungen (Locations, Stadtteile ..)
    - Teilnehmerzahl

Auswertungen der Datenqualität der Quelldaten im Netz

- Unvollständige bzw. fehlende Informationen (z.B. bei Adressen von Anbietern)
- Veraltete Informationen (z.B. nicht-existente oder umbenannte Gaststätten)
- Inkonsistente (widersprüchliche) Informationen (z.B. widersprüchliche Veranstaltungs-bezeichnungen aus verschiedenen Quellen)
- Offenkundige Fehler in den Einträgen

Fallstudie 2: HTWK-Data Warehouse

Aufbau eines Data Warehouse mit öffentlich zugänglichen Informationen über die HTWK Leipzig. Im Data Warehouse sollten Informationen zu finden sein über Professoren & Mitarbeiter der Hochschule, deren Lehrangebote und Forschungsergebnisse:

- Professoren und Mitarbeiter
    - Adressinformationen (z.B. E-Mail, Homepage, FB, Telefon, Büro, Sprechzeiten)
    - Berufungsgebiet / Lehrgebiete
    - Funktionen an FB / Hochschule
- Lehrveranstaltungen
    - Persönlicher Stundenplan (Lehrveranstaltungen und Termine)
    - Studentische Bewertungen (meinprof.de)
- Forschung
    - Forschungsthemen
    - Bisher betreute Diplomarbeiten
    - Veröffentlichungen

Mögliche Auswertungen

- Anzahl und Art der Veröffentlichungen (Google Scholar, für Informatik: DBLP Trier)
- Statistik über Lehrauslastung
    - Bedienbeziehungen (z.B. Welcher FB bedient welchen anderen FB?)
    - Auswertungen pro Professor / Fachbereich u.a.
- Statistik über Lehrangebote
- Statistik über betreute Diplomarbeiten (sofern Informationen online)
- Mögliche Auswertedimensionen: Fachbereich, Fachgebiet, Professor

Auswertungen der Datenqualität der Quelldaten im Netz

- Unvollständige bzw. fehlende Informationen (z.B. bei Diplomthemen)
- Veraltete Informationen (z.B. ausgeschiedene Professoren)
- Inkonsistente (Widersprüchliche) Informationen (z.B. Zugehörigkeit zu FB, akademischer Grad)
- Fehler in den Einträgen (z.B. bei Büro-Adresse)

.

Es können auf besonderen Wunsch auch andere Fallbeispiele bearbeitet werden. Diese müssen jedoch zu Beginn des Projekts abgesprochen werden.