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

Erläuterung zum Projekt zur Lehrveranstaltung “Informationssysteme“

Data Warehouse = physische Datenbank, die eine *integrierte* Sicht auf beliebige Daten ermöglicht.

Referenzarchitektur – siehe Folie

Datenquellen

Enthalten alle für das DWH relevanten Daten, Ausgangspunkt einer datenflussorientierten Betrachtung der DWH-Architektur

Intern: operative Systeme der eigenen Organisation (z.B. HIS, S-PLUS) – nicht für studentische Benutzer

Extern: Internet-Quellen (HTML-Seiten, Webservices), aber auch Intranet-Angebote (z.B. Stundenplan, Telefonverzeichnis)

Auswahl der Quellen? Qualität, Verfügbarkeit, Preis

Back-End-Werkzeuge:

ETL: Chickenfoot, FireBug

Oracle Warehouse Builder

Datenqualitätssicherung: wichtige Aufgabe im Back-End-Bereich

Dimensionen der Datenqualität – siehe Vorlesung

Operational Data Store: Datenbank, die als physischer Zwischenspeicher genutzt wird (z.B. nach dem Data Cleansing einzelner Quellen, evtl. noch ohne Zeitbezug)

Monitoring: Entdeckung von Datenänderungen

Veränderungen inkrementell ins DWH propagieren

Verschiedene Monitoringstrategien, abhängig von der benötigten Aktualität der Daten

Bei niedriger Änderungsfrequenz: periodische Auswertung von Datenquellen, Vergleich

aufeinanderfolgender Snapshots (Delta-Berechnung) identifiziert Änderungen – typische

Anwendung bei Legacy-Systemen ohne eingebaute Monitoring-Komponente und bei „Black-Box-Sicht“ auf Systeme

Extraktion:

abhängig von Monitoring-Strategie, Identifikation der geänderten Tupel in Dateien mit einer bestimmten Struktur (schwierig bei Dateien als Datenquelle)

Zeitstempel zum Markieren der Extraktions-/Ladezeiten empfehlenswert!

Transformation:

Ausgangssituation: mehrere heterogene Datenquellen (Daten mit unterschiedlichen und überschneidenden Inhalten in unterschiedlichen Repräsentationsformen)

Vor dem Laden: Überführen in einheitliche Repräsentation (Data Migration)

Beispiel: Excel-Datei vs. XML-Datei

Speicherung der transformierten Daten im ODS, evtl. mit Historisierung inkrementeller Änderungen (grundsätzlich kein Überschreiben vorgesehen, sondern zusätzlicher Datensatz ins DWH ggf. mit einem Zeitstempel versehen)

Voraussetzung für Datentransformation? Transformation von Datenstrukturen bzw. Schemata (relationales DB-Schema, WSDL, XML-Schema, DTD)

Laden:

Übertragung konsolidierter Daten aus dem ODS in die analyseorientierten Strukturen des DWH

Initiales Laden offline, ansonsten parallel zum laufenden Betrieb des DWH

Archiv-Datenbank

Optional für definierte „Entsorgung“ veralteter Daten

Front-End-Werkzeuge: Reporting, OLAP, Data Mining

Für uns: OLAP zur interaktiven Datenanalyse, multidimensionale Sicht auf die Daten

Anwender kann in Abhängigkeit von der aktuellen Anfrage eine Zusammenstellung der Daten hinsichtlich Verdichtungsgrad und Präsentation vornehmen (auf individuelle Bedürfnisse zugeschnitten – flexibler als Reporting)

Sonstige Werkzeuge:

DWS-Manager (für Administrator)

Metadaten-Repository

Administrative Metadaten: Schemadaten, Infos über Quell- und Zielsysteme,

Datenabhängigkeiten / Transformationsregeln

Operative Metadaten: Systemstatistiken, Logs u.ä. – für uns bei Entwurf nicht relevant

Domänenspezifische Metadaten:

- Informationsmodelle und konzeptuelle Schemata (implementierungsunabhängige Dokumentation)
- Begriffswerke (Taxonomie, Vokabular)
- Organisationsstrukturen und Geschäftsprozesse
- Angaben über die Datenqualität
- Konzeptionelle Beschreibungen von Berichten, Anfragen, Kennzahlen