Vorlesung
Data-Warehouse-Technologien

Wintersemester 2002/2003

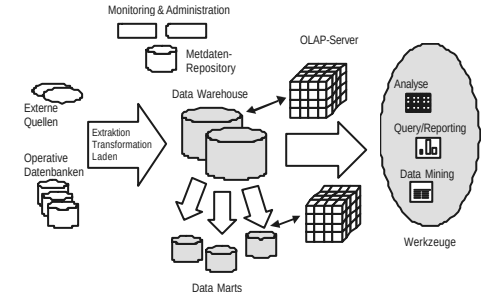
Kai-Uwe Sattler

kus@iti.cs.uni-magdeburg.de
http://wwwiti.cs.uni-magdeburg.de/~sattler/hal/dw.html

Einführung

- Gegenstand der Vorlesung
  - Data Warehouse: Sammlung von Technologien zur Unterstützung von Entscheidungsprozessen
  - Herausforderung an Datenbanktechnologien
    - Datenvolumen (effiziente Speicherung und Verwaltung, Anfragebearbeitung)
    - Datenmodellierung (Zeitbezug, mehrere Dimensionen)
    - Integration heterogener Datenbestände
- Schwerpunkt
  - Datenbanktechniken von Data Warehouses

Überblick



Betriebswirtschaftliche Anwendungen

- Informationsbereitstellung
  - Daten und Informationen als Grundlage einer erfolgreichen Abwicklung von Geschäftsprozessen (z.B. Kennzahlen)
  - Anwender: Manager, Abteilungsleiter, Fachkräfte
  - Formen der Bereitstellung
    - Query-Ansätze: frei definierbare Anfragen und Berichte
    - Reporting: Zugriff auf vordefinierte Berichte
    - Redaktionell aufbereitete, personalisierte Informationen

Betriebswirtschaftliche Anwendungen

- Analyse
  - Detaillierte Analyse der Daten zur Untersuchung von Abweichungen oder Auffälligkeiten
  - Anwender: Spezialisten (z.B. Controlling, Marketing)
- Planung
  - Unterstützung durch explorative Datenanalyse
  - Aggregrierung von Einzelplänen
- Kampagnenmanagement
  - Unterstützung strategischer Kampagnen
  - Kundenanalyse, Risikoanalyse

Wissenschaftliche und Technische Anwendungen

- Wissenschaftliche Anwendungen
  - *Statistical und Scientific Databases* ? technische Wurzeln des DW
  - Beispiel: Projekt Earth Observing System (Klima- und Umweltforschung)
    - täglich ca. 1,9 TB meteorologischer Daten
    - Aufbereitung und Analyse (statistisch, Data Mining)
- Technische Anwendungen
  - Öffentlicher Bereich: DW mit Umwelt- oder geographischen Daten (z.B. Wasseranalysen)

Einsatzbeispiel

- Wal▸Mart (www.wal-mart.com)
- Marktführer im amerikanischen Einzelhandel
- Unternehmensweites Data Warehouse
  - Größe: ca. 25 TB
  - Täglich bis zu 20.000 DW-Anfragen
  - Hoher Detaillierungsgrad (tägliche Auswertung von Artikelumsätzen, Lagerbestand, Kundenverhalten)
  - Basis für Warenkorbanalyse, Kundenklassifizierung, ...

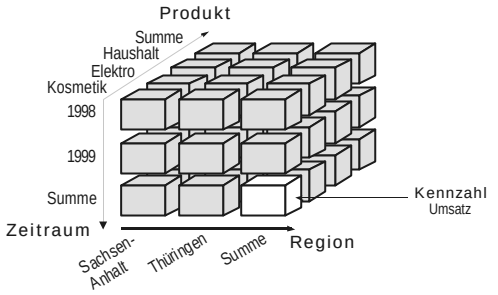
Fragestellungen und Aufgaben (Bsp.)

- Überprüfung des Warensortiments zur Erkennung von Ladenhütern oder Verkaufsschlagern
- Standortanalyse zur Einschätzung der Rentabilität von Niederlassungen
- Untersuchung der Wirksamkeit von Marketing-Aktionen
- Auswertung von Kundenbefragungen, Reklamationen bzgl. Bestimmer Produkte etc.
- Analyse des Lagerbestandes
- Warenkorbanalyse mit Hilfe der Kassenbons

Beispiel einer Anfrage

Welche Umsätze sind in den Jahren 1998 und 1999 in den Abteilungen Kosmetik, Elektro und Haushaltswaren in den Bundesländern Sachsen-Anhalt und Thüringen angefallen ?

Ergebnis (Würfel)

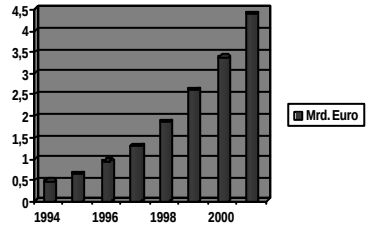


Ergebnis (Bericht)

Umsatz		Kosmetik	Elektro	Haushalt	SUMME
1998	Sachsen-Anhalt	45	123	17	185
	Thüringen	43	131	21	195
	SUMME	88	254	38	380
1999	Sachsen-Anhalt	47	131	19	197
	Thüringen	40	136	20	196
	SUMME	87	267	39	393
SUMME		175	521	77	773

Marktentwicklung

☞ Marktgröße: Data Warehouse und OLAP (Quelle: OLAP Report OnLine www.olapreport.com)



Aspekte von Data Warehouses

- **Integration**
  - Vereinigung von Daten aus verschiedenen, meist heterogenen Quellen
  - Überwindung der Heterogenität auf verschiedenen Ebenen (System, Schema, Daten)
- **Analyse**
  - Bereitstellung der Daten in einer vom Anwender gewünschten Form (bezogen auf Entscheidungsgebiet)
  - erfordert Vorauswahl, Zeitbezug, Aggregation

Abgrenzung zu OLTP

- Klassische operative Informationssysteme
  - ? *Online Transactional Processing (OLTP)*
  - Erfassung und Verwaltung von Daten
  - Verabeitung unter Verantwortung der jeweiligen Abteilung
  - Transaktionale Verarbeitung: kurze Lese-/ Schreibzugriffe auf wenige Datensätze
- Data Warehouse
  - Analyse im Mittelpunkt
  - lange Lesetransaktionen auf vielen Datensätzen
  - Integration, Konsolidierung und Aggregation der Daten

Abgrenzung zu OLTP: Anfragen

Anfrage	**transaktional**	**analytisch**
Fokus	Lesen, Schreiben, Modifizieren, Löschen	Lesen, periodisches Hinzufügen
Transaktionsdauer und –typ	kurze Lese-/ Schreibtransaktionen	lange Lesetransaktionen
Anfragestruktur	einfach strukturiert	komplex
Datenvolumen einer Anfrage	wenige Datensätze	viele Datensätze
Datenmodell	anfrageflexibel	analysebezogen

Abgrenzung zu OLTP: Daten

Daten	transaktional	analytisch
Datenquellen	meist eine	mehrere
Eigenschaften	nicht abgeleitet, zeitaktuell, autonom, dynamisch	abgeleitet/konsolidiert, nicht zeitaktuell, integriert, stabil
Datenvolumen	MByte ... GByte	GByte ... TByte
Zugriffe	Einzeltupelzugriff	Tabellenzugriff

Abgrenzung zu OLTP: Anwender

Anwender	transaktional	analytisch
Anwendertyp	Ein-/Ausgabe durch Angestellte oder Applikationssoftware	Manager, Controller Analyst
Anwenderzahl	sehr viele	wenige (bis einige hundert)
Antwortzeit	ms ... sec	sec ... min

Data Warehouse: Begriff

A Data Warehouse is a subject-oriented, integrated, non-volatile, and time variant collection of data in support of managements decisions.

(W.H. Inmon 1996)

Data Warehouse: Charakteristika

- Fachorientierung (subject-oriented):
  - Zweck des Systems ist nicht Erfüllung einer Aufgabe (z.B. Personaldatenverwaltung), sondern Modellierung eines spezifischen Anwendungsziels
- Integrierte Datenbasis (integrated):
  - Verarbeitung von Daten aus mehreren verschiedenen Datenquellen (intern und extern)
- Nicht-flüchtige Datenbasis (non-volatile):
  - stabile, persistente Datenbasis
  - Daten im DW werden nicht mehr entfernt oder geändert
- Historische Daten (time-variant):
  - Vergleich der Daten über Zeit möglich (Zeitreihenanalyse)
  - Speicherung über längeren Zeitraum

Weitere Begriffe

- Data Warehousing
  - Data-Warehouse-Prozess, d.h. alle Schritte der Datenbeschaffung (Extraktion, Transformation, Laden), des Speicherns und der Analyse
- Data Mart
  - externe (Teil-)Sicht auf das Data Warehouse
  - durch Kopieren
  - anwendungsbereichsspezifisch
- OLAP (*Online Analytical Processing*)
  - explorative, interaktive Analyse auf Basis des konzeptuellen Datenmodells

Trennung operativer und analytischer Systeme

- Gründe
  - Antwortzeitverhalten: Analyse auf operativen Quelldatensystemen ? schlechte Performance
  - Langfristige Speicherung der Daten ? Zeitreihenanalyse
  - Zugriff auf Daten unabhängig von operativen Datenquellen (Verfügbarkeit, Integrationsproblematik)
  - Vereinheitlichung des Datenformats im DW
  - Gewährleistung der Datenqualität im DW

Historie

- Wurzeln
  - 60er Jahre: ExecutiveInformation Systems (EIS)
    - qualitative Informationsversorgung von Entscheidern
    - kleine, verdichtete Extrakte der operativen Datenbestände
    - Aufbereitung in Form statischer Berichte
    - Mainframe
  - 80er Jahre: Management Information Systems (MIS)
    - meist statische Berichtsgeneratoren
    - Einführung von Hierarchieebenen für Auswertung von Kennzahlen (Roll-Up, Drill-Down)
    - Client-Server-Architekturen, GUI (Windows, Apple)

Historie

- 1992: Einführung des Data-Warehouse-Konzeptes durch W.H. Inmon
  - redundante Haltung von Daten, losgelöst von Quellsystemen
  - Beschränkung der Daten auf Analysezweck
- 1993: Definition des Begriffs OLAP durch E.F. Codd
  - Dynamische, multidimensionale Analyse
- Weitere Einflußgebiete
  - Verbreitung geschäftsprozeßorientierter Transaktionssysteme (SAP R/3) ? Bereitstellung von entscheidungsrelevanten Informationen
  - Data Mining
  - WWW (Web-enabled Data Warehouse etc.)

Vorlesung: Zielstellungen

- Vermittlung von Kenntnissen zu *Datenbanktechniken* für Aufbau und Implementierung von Data Warehouses
- Anwendung bekannter DB-Techniken (siehe Vorlesung „Datenbanken I“)
  - Datenmodellierung, Anfragesprachen und -verarbeitung
- DW-spezifische Techniken
  - multidimensionale Datenmodellierung
  - spezielle Anfragetechniken
  - Indexstrukturen
  - materialisierte Sichten

DW-Architektur

- Komponenten von DW und deren Aufgaben
- Datenbanken
  - Datenquellen: Herkunftsort der Daten
  - Arbeitsbereich: temporäre Datenbank für Transformation
  - Data Warehouse: physische Datenbank für Analyse
  - Repository: Datenbank mit Metadaten

DW-Architektur

- Komponenten
  - Data-Warehouse-Manager: zentrale Kontrolle und Steuerung
  - Monitore: Überwachung der Quellen auf Veränderungen
  - Extraktoren: Selektion und Transport der Daten aus Quellen in Arbeitsbereich
  - Transformatoren: Vereinheitlichung und Bereinigung der Daten
  - Ladekomponenten: Laden der transformierten Daten in das DW
  - Analysekomponenten: Analyse und Präsentation der Daten

Multidimensionales Datenmodell

- Datenmodell zur Unterstützung der Analyse
  - Fakten und Dimensionen
  - Klassifikationsschema
  - Würfel
  - Operationen: Pivotierung, Roll-Up, Drill-Down, Drill-Across, Slice und Dice
- Notationen zur konzeptuellen Modellierung
- Relationale Umsetzung
  - Star-Schema, Snowflake-Schema
- Multidimensionale Speicherung

Anfrageverarbeitung und -optimierung

- Gruppierung und Aggregation
  - Supergroups, CUBE
- Star-Joins
- Optimierungsaspekte
  - Histogramme, Sampling
- Mehrdimensionale Erweiterungen von Anfragesprachen
  - MDX

Index- und Speicherungsstrukturen

- Klassifikation
- Wiederholung: B-Baum und B*-Baum
- Mehrdimensionale Indexstrukturen
  - R-Baum
  - UB-Baum
- Bitmap-Index
- Vergleich
- Multidimensionale Speicherung

Materialisierte Sichten

- Materialisierte Sicht (engl. *materialized view*): vorab berechneter Ausschnitt aus einer Faktentabelle
- Verwendung: Anfrageersetzung
  - *generalized projection*
- Auswahl: Bestimmung der redundant gehaltenen Daten
  - statische vs. dynamische Auswahlverfahren
  - Semantisches Caching
- Wartung und Aktualisierung

Metadaten und Datenqualität

- Metadatenmanagement
- Metadaten-Repository
- Standards für Metadaten
- Aspekte der Datenqualität

OLAP und Data Mining

- OLAP
  - Anforderungen
  - OLAP-Operationen
  - OLAP-Werkzeuge
- Data-Mining-Technikem
  - Klassifikation, Assoziationsregeln, Clustering

Aufbau und Betrieb von DW

- Aufbau
  - Phasen
  - Regeln für den Aufbau
- Betrieb
  - Administration
  - Sicherheitsmanagement
  - Performance-Tuning

Tools und Systeme

- Oracle
  - ETL-Werkzeuge
  - DW-Erweiterungen für Oracle-Server
- IBM DB2
  - OLAP-Server
- ...

Literatur

- Lehner: *Datenbanktechnologie für Data-Warehouse-Systeme*, dpunkt.verlag, 2002
- Bauer, Günzel (Hrg.): *Data Warehouse – Architektur, Entwicklung, Anwendung*; dpunkt.verlag, 2000
- Jarke, Lenzerini, Vassiliou, Vassiliadis: *Fundamentals of Data Warehouses*; Springer Verlag, 2000
- Kurz: *Data Warehousing: Enabling Technology*; MITP, 1999