Near Real-Time Data Integration
VO1: Einführung in Data Warehousing

Konzepte und Techniken von Data Warehouses

Robert M. Bruckner
bruckner@ifs.tuwien.ac.at

Institut für Softwaretechnik &
Interaktive Systeme
Technische Universität Wien

Data Warehouse - Agenda

Teil 1:

- Einführung
- DWH Charakteristik
- DWH Architektur
- DWH Typen und Komponenten

Teil 2:

- Einführung in die Teradata Datenbank
- „Traditionelles“ Laden von Daten
  - FastLoad
  - MultiLoad

DWH Architektur

- Einführung / Überblick
- Abgrenzung OLTP und DWH/OLAP
- Data Warehouse Historie
- Architekturen und Struktur eines Data Warehouses
- Komponenten eines Data Warehouses
  - Ladevorgang
  - Analysevorgang
  - Metadaten
- Data Warehouse Projekt

Definition nach Inmon

„A Data Warehouse is a subject-oriented, integrated, non-volatile, time-variant collection of data organized to support management needs“

- **themenorientiert**
  alle Daten über ein Subjekt (z.B. Kunde) zentral gespeichert und nicht in einzelnen Applikationen “versteckt”
- **integriert**
  Daten aus heterogenen Systemen und Formaten einheitlich in gemeinsamen Schema gespeichert und von Inkonsistenzen bereinigt
- **historisch**
  Daten (Snapshots) über lange Zeit (5-10 Jahre) gespeichert, um Zeitanalysen vornehmen zu können
- **nicht flüchtig**
  Das DWH wird periodisch aktualisiert; einmal eingebrachte Daten dürfen nicht verändert werden

Motivation Data Warehouses

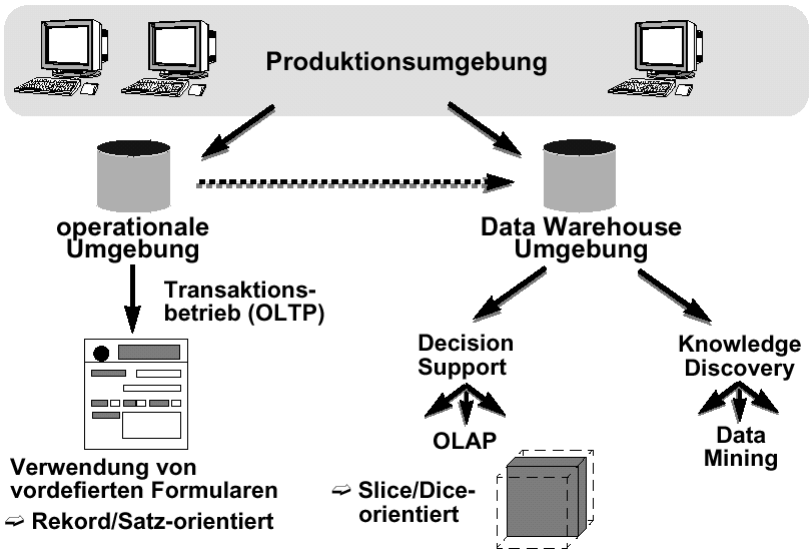
- Heterogene Daten integrieren und anreichern, um sie einer homogenen Analyse zuzuführen, die Informationen für (strategische) Entscheidungen liefert.

- Wettbewerb
  - Globale Märkte, schnelle Veränderungen
  - Information wird immer wichtiger für Unternehmen ⇔ Informationsüberflutung
  - Entscheidungsträger benötigen last-minute hoch-aggregierte Informationen
  - Kunden flexibler
  - → Kürzere decision times basierend auf globaler (integrierter) Information
  - → Mehr Wissen über den Kunden (CRM – Customer Relationship Management)

Motivation Data Warehouses (2)

- Technologie
  - Effiziente Datenbanktechnologien verfügbar
  - Verteilte Architekturen
  - Hardware / Storage immer leistungsfähiger und billiger
  - User-Interfaces werden besser → Nutzung der IT-Infrastruktur durch mehr Personen
- Trends
  - OLAP (On-Line Analytical Processing)
  - Intranet
  - Internet / WWW
  - Data Mining
  - Knowledge Discovery in Databases (KDD)

Beispiel für Umgebung eines DWH



Weitere Definitionen für DWH

A Data Warehouse is:

- a method for combining information from a number of different systems
- a central access point for an organization's data
- populated by data from otherwise incompatible information systems
- presented in a business relevant format
- "read-only" information, which means you can't affect the information sources
- enables reporting across many of these host applications

(Pandora, Swinburn University)

Weitere Definitionen für DWH (2)

A Data Warehouse is:

- a complete and consistent store of data
- obtained from a variety of sources
- made available to end users
- in a way they can understand
- and use in a business context

(Paul Lucas, IBM)

Definition nach Inmon

A Data Warehouse is a ...

- subject-oriented,
- integrated,
- time-variant,
- nonvolatile

... collection of data in support of management's decision making process.

(W. H. Inmon)

DWH Charakteristik

- Design für analytische Aufgaben (d.h. komplexe Abfragen)
- Relativ kleine Anzahl an Benutzern
- Lese-intensiver Zugriff
- Periodische Updates (Hinzufügen von Daten)
- Historisierung der Daten
- Abfragen liefern umfangreiche Antwortmengen, häufig Full-Table Scans kombiniert mit Joins von vielen Tabellen
- Globale Sicht

Datenhaltung – Divergierende Ziele

OLTP (Online Transactional Processing)

- Technologie: Relationale DBMS
- Transaktionsorientiert
- Möglichst redundanzfrei, hochnormalisiert
- Fokus auf aktuelle Daten
- Abfragen liefern kleine Ergebnismengen
- Transaktionen: Integrität, Security, Concurrency, Locking
- Transaktionsdurchsatz
- Hoch-Verfügbarkeit (24x7)
- Workload auf gleichmäßig hohem Level

Data Warehouse / OLAP (Online Analytical Processing)

- RDBMS, MDBMS (multidimensionale DBMS)
- Teilweise hoch-redundant
- Vorberechnete Aggregationen
- Historische Daten / Analysen
- Analysen sehr komplex
- Gesamtintegration verschiedenster Systeme
- Themenorientiert
- Sehr hohe Datenvolumina
- Geringere Anforderungen an Verfügbarkeit
- Spitzenlasten, ad-hoc Abfragen

Datenhaltung – Divergierende Ziele (2)

OLTP (Online Transactional Processing)

- Hohe Zahl an parallelen Transaktionen
- Updates sehr häufig
- Zeitliche „Inkonsistenzen“: d.h. durch Daten-Updates kann es passieren, dass dieselbe Abfrage unterschiedliche Resultate zu unterschiedlichen Zeitpunkten liefert.
  Bsp: Relation Part-Supplier
- Abfragen sind oftmals vorher bekannt und optimiert

Data Warehouse / OLAP (Online Analytical Processing)

- Antwortzeit ist nicht der primäre Fokus für Abfragen
- Keine Updates (traditionell) oder Deletes von DWH Daten
- Historisierung der Daten → im allgemeinen zeitlich konsistente Ergebnisse
- Neue Daten aus den Vorsystemen (DB, Legacy Systeme, sonstige Datenquellen, auch externe Datenquellen – z.B. Ortsdatenbank) werden zu vordefinierten Zeitpunkten (traditionell) integriert
- Ad-hoc Abfragen

Abgrenzung OLTP – DWH/OLAP

Operationale Systeme

- Schnelle Antwortzeit
- Applikationsorientiert
- Aktuelle Daten
- Daten in hoher Granularität (Detailierungsgrad)
- Häufige Updates
- Viele (kurze) Transaktionen zur Abwicklung des Tagesgeschäft
- Überwiegend konstante, hohe Last
- Oftmals isolierte Systeme

Informationssysteme (DWH/OLAP)

- Hohes Datenvolumen
- Themenorientiert (z.B. Kundensicht)
- Historische Daten
- Teilweise Vor-aggregierte Daten, konsolidierte Daten
- Keine Updates
- Komplexe Analysen für Entscheidungsunterstützung
- Spitzenlasten und geringe Last
- Integration von heterogenen Vorsystemen

Abgrenzung OLTP – DWH/OLAP

Charakteristika	OLTP	DWH/OLAP	
Benutzertyp	Angestellter, IT-Professional	Manager, Decision Support	Benutzer Anfragen
Benutzeranzahl	Hoch (Tausende), Concurrency	Wenige Benutzer	
Antwortzeit	Sekunden	Sekunden – Minuten	
Anwendung	Verwaltung, operatives Geschäft	Analyse und Entscheidungsunterstützung	
Fokus	Dateneingabe	Informationsgewinn	
Auswertungsmethode	Datensatz-orientiert	Multidimensional	
Transaktionsart	Kurze Lese-/Schreibtransaktionen	Lange Lesetransaktionen	
Funktion	Tägliche Operationen	Wissensgenerierung	
Einheit	Einfache Transaktionen	Komplexe Anfragen	
Ergebnismenge	Klein	Sehr groß	
Anfragetyp	Strukturiert, vordefiniert	Ad-hoc	

Abgrenzung OLTP – DWH/OLAP

Charakteristika	OLTP	DWH/OLAP	
Datenverwaltungsziel	Transaktionale Konsistenzerhaltung	Zeitbasierte Versionierung	Daten
Daten	Aktuell, atomar, isoliert, dynamisch	Historisch, konsolidiert, stabil, aggregiert, integriert, multidimensional	
Datenvolumen	MB – GB	GB – TB	
Design	Anwendungsorientiert (ER basierend), hochnormalisiert	Subject-oriented (Star-/Snowflake), teilweise hochnormalisiert	
Datensicht	Detailliert, komplex, relational	Einfach, summiert, multidimensional	
Typische Operationen	Index/Hashzugriff auf Primärschlüssel	(Full-Table) Scans	
Anpassungsfähigkeit	Begrenzt	Anwendungsabhängig	

Architektur eines DWH

“Data Warehouse is an environment, not a product”
(Berson/ Smith)

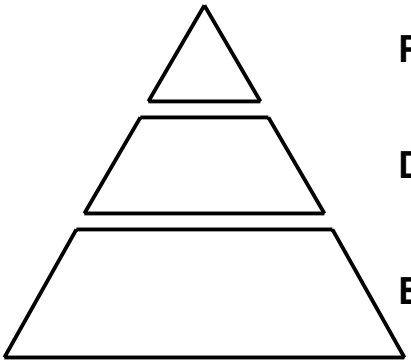
Architektonisches Konstrukt eines Informationssystems

- Systeminfrastruktur: Hardware, Software, Netzwerk, DBMS
- “Sensibilität” für die Daten: Verständnis für die vorhandenen Daten (auch Datenqualität), die geladen werden könnten
- Datenaquirierung: Laden der Daten aus den unterschiedlichen Quellen
- Datenverteilung: Speicherung der Daten im Warehouse
- Benutzeranalyse: Verwendung der gespeicherten Daten
- Metadaten: Daten über Daten

Gründe für ein separates DWH

- Konzeptionelle Differenzen im Datenbestand und dessen Nutzung
- Performance
  - OLTP optimiert für kurze Transaktionen und bekannte Lastprofile
  - komplexe OLAP-Anfragen degradieren die Performanz von Transaktionen des operationalen Betriebs
  - spezieller physischer und logischer Datenbankentwurf für multidimensionale Sichten und Anfragen notwendig
  - Transaktionseigenschaften (ACID) nicht wichtig
- Funktionalität
  - für OLAP-Anfragen werden historische Daten benötigt, die in OLTP-Systemen typischerweise nicht vorliegen
  - Konsolidierung (Integration, Bereinigung und Aggregation) von Daten aus heterogenen Datenquellen
- Sicherheit

DWH Schichtarchitektur



Presentation Layer

DWH / OLAP Layer

Extraction Layer

DWH Schichtarchitektur: Extraktion

Laden des Data Warehouse

- Lesen von Daten aus beliebigen Quellen (ODBC, Excel, Internet, etc.)
  - Altsysteme oft undokumentiert und schwer zugänglich
- Entfernen von semantischen Inkonsistenzen (Integritätsprüfungen)
  - Außerhalb der Quellen mit überprüfbaren Qualitätssicherungsmethoden
- Effizienz des Ladevorgangs: “Kampf dem Ladefenster” (falls es eines gibt ...)
  - inkrementelles Laden; Erstellen/Aktualisieren von Aggregationen während des Ladeprozesses
  - online oder offline Laden? (Sperren von Datenpartitionen, blockieren von Anfragen)
  - paralleles Laden

DWH Schichtarchitektur: DWH / OLAP

Data Warehouse Datenhaltung

- Analysen benötigen oft Daten auf Rohdatenebene (feinste Granularität)
  - Verwaltung von sehr großen Datenvolumina (GB → TB)
- Datenorganisation innerhalb des Data Warehouses:
  - anwendungsorientierte Aufteilung nach Dimensionen (nach der Zeit, nach Produkten, etc.)
  - zeitbasierte Datenverwaltung ist unbedingt notwendig (Ladevorgänge sind häufig zeitorientiert)
- Verfügbarkeitsanforderung steigen (Trend 24x7 System)
  - DWH ist nicht mehr nur ein 'add-on' oder ein 'Spielzeug' für Entscheidungsträger
  - DWH wird dazu verwendet, Geld zu "machen" (break-even nach 3 Jahren; MetaGroup)

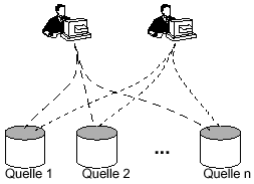
DWH Schichtarchitektur: Presentation

Analyse / Decision Support / Presentation

- "We are drowning in data, but starving for knowledge"
- Multidimensionalität als intuitiver Ansatz
  - effizienten Zugriff auf Datenwürfel bereitstellen
  - Unterstützung von verbreiteten Tabellenkalkulationen
- Typische Anfragen:
  - Wie haben sich dieses Jahr die Verkaufszahlen nach Verkaufskanäle und Produktgruppen entwickelt (im Vergleich zum letzten Jahr)?
  - Welche Geschäfte machen 80 Prozent des Umsatzes einer bestimmten Marke?
  - Wer sind die Hauptkonkurrenten eines Elektronikherstellers?
  - Distributionsüberschneidungen: Aufdecken von Korrelationen zwischen zwei Marken A und B

Historie / Grund für DWH Schichtarchitektur

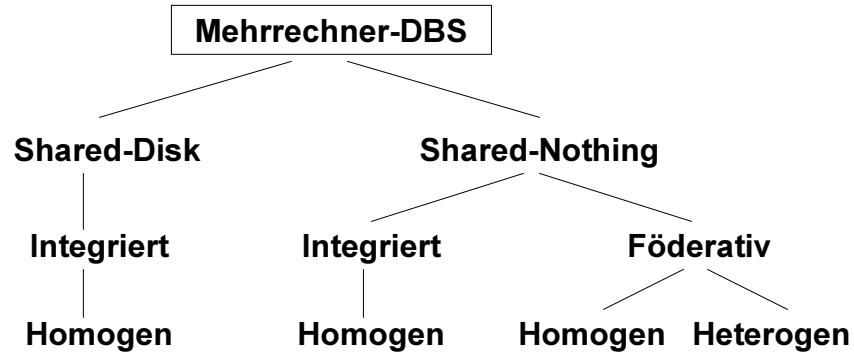
- Ziel: Integration heterogener Datenbanken
- Leichter und effizienter Zugriff auf integrierte Informationen aus vielen verschiedenen heterogenen, autonomen und verteilten Informationsquellen



- Ziel und Lösungsansätze schon länger vorhanden: verteilte DBS, föderative DBS
- Analytische Queries

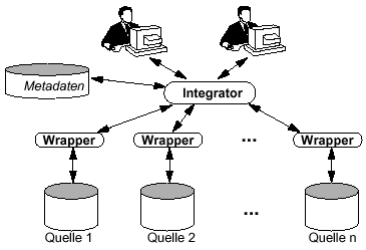
  Online Analytical Processing (OLAP); Notwendigkeit von Datenredundanz (Aggregationen) zur Performanzsteigerung

Alternative: Integrierte / Föderative DBMS ?



Alternative: Integrierte / Föderative DBMS ?

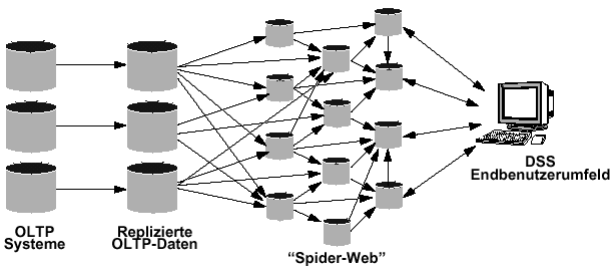
- Anfragegesteuert, “on demand”, “virtueller Ansatz”



- Vorteile:
  - Bei sich sehr schnell ändernden Daten
  - Bei unvorhersehbaren Anfragetypen und Lastprofilen
- Nachteile:
  - lange Anfragelaufzeiten wegen Anfrageumsetzung (Integrator/Wrapper)
  - für häufige Anfragen ineffizient
  - konkurriert mit operationalem Betrieb auf den einzelnen Quellsystemen

Alternative: MIS / EIS ?

- MIS/EIS: Management / Executive Information System
- Nur vordefinierte Anfragen bzw. Berichtsgenerierung
- Separate, konsolidierte Datenbasis fehlt – oft Duldung von Inkonsistenzen
- DWH kann auf ein MIS / EIS aufgesetzt werden



DWH Ansatz

- Information a priori integriert und im Data Warehouse gespeichert

- Vorteile:
  - relativ kurze Anfragelaufzeiten
  - konkurriert nicht mit operationalem Betrieb
  - Externe Datenquellen leichter integrierbar
- Nachteil:
  - Redundanz und geringere Aktualität der Daten
    → neue Entwicklungen: z.B. (Near) Real-Time DWHs

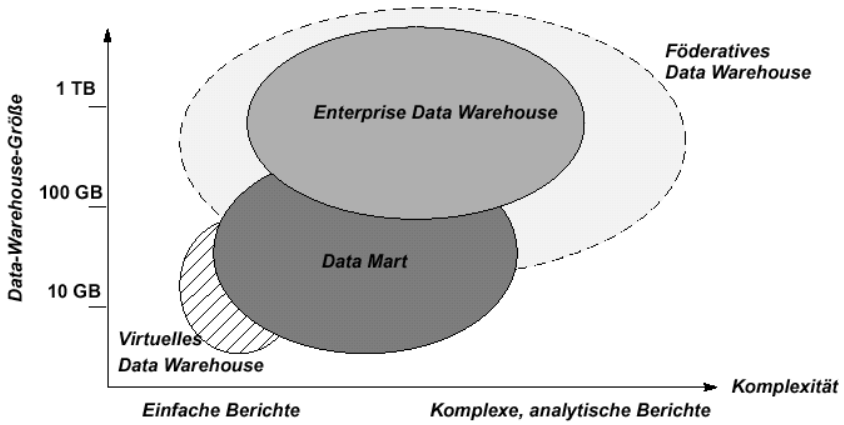
Historische Entwicklung

- 60er Jahre: Berichte im Batch-Betrieb
  - Information ist schwer zu finden und zu analysieren (Bänder!)
  - Nicht flexibel und teuer; jede Anfrage / jeder Bericht muß neu programmiert werden
- 70er Jahre: Terminal-basierte DSS und EIS
  - Nicht in die operationale Umgebung integriert
  - Immer noch unflexibel
- 80er Jahre: Desktop Datenzugriff und Analyseprogramme
  - Anfragetools, Spreadsheets, GUIs
  - Leichte Bedienung, jedoch Zugriff beschränkt auf operationale Datenbanken
- 90er Jahre:
  - Data Warehousing mit integrierten OLAP-Servern und Tools
- Heute: DWH als Mission-Critical System

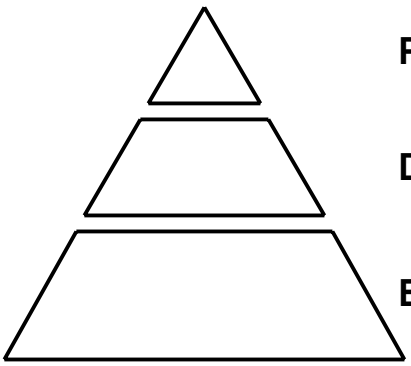
Verwendungszweck / Evolution von DWHs

- **Reporting**
  Was ist passiert?
- **Analyse**
  Warum ist etwas passiert?
- **Vorhersage / Prediction**
  Was wird passieren?
- **Operationalisierung**
  Was ereignet sich jetzt gerade?
- **Active DWH**
  Was soll sich ereignen?
- **Closed Loop Decision Support**
  Zielbasierter Regelansatz: Welche Effekte werden Entscheidungen haben?

DWH Typen



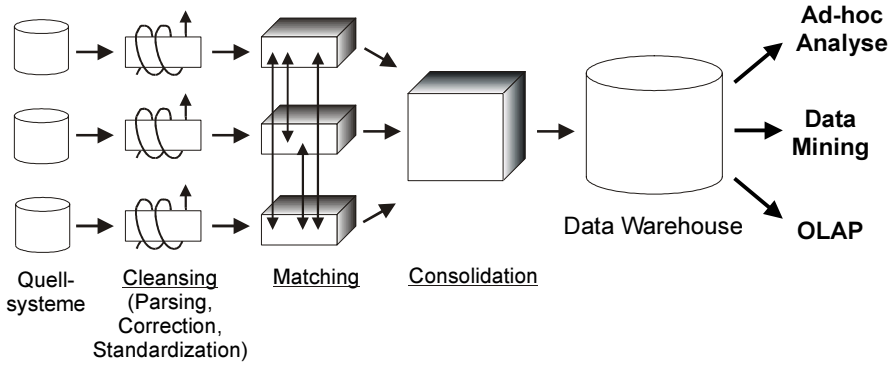
DWH Architektur im Detail



DWH: ETL / Datenqualität

ETL: Extract - Transform – Load

Datenqualität begrenzt die Verwend-/Verwertbarkeit der Daten



DWH: ETL / Datenqualität

- Methoden der Datenqualitätsanalyse
  - Sicht des Datenproduzenten: intuitiver (projektabhängiger) Ansatz und theoretischer Ansatz (Suche nach methodischen Fehlern)
  - Sicht des Datenkonsumenten: empirischer Ansatz des Endverbrauchers
- DQ-Indikatoren:
  - Daten sind präzise
  - Daten sind in einem bestimmten Datentyp gespeichert
  - Daten sind konsistent
  - sinnvolles Datenbankdesign
  - keine Redundanz
  - Daten folgen den Business Regeln
  - Domains werden eingehalten
  - Zeitrichtig (timeliness)
  - …

DWH: ETL / Datenqualität

- DQ-Indikatoren (Fortsetzung):
  - gut verständlich
  - Daten sind integriert
  - Daten passen zum Geschäftsfall
  - Benutzer ist mit den Daten zufrieden
  - Daten sind komplett
  - keine doppelten Felder
  - keine Datenanomalien
- Daten sind normalerweise sehr schlecht
  → Erkennung ist notwendig
  - automatische Ausreißererkennung
  - Benutzer, die Anomalien erkennen oder erahnen
  - Datenalter
  - Programmabbrüche

DWH: ETL / Datenintegration

- Überlegung:
  - Laden von Daten aus unabhängigen Systemen,
  - inkompatiblen Datenformaten,
  - redundanten Daten
  - und aus unterschiedlichen Begriffswelten
- Ziel:
  - Integration heißt Datenmaterial und Verwendbarkeit verbessern
  - keine einheitliche Technologie, Schnittstellen, Datenformate
  - DWH braucht aber genau diese Vereinigung (ca. 2/3 des Aufwandes beim Aufbau eines DWH)
- Vorgehensweise:
  - Schemaintegration
  - Laden der Daten

DWH: ETL / Schemaintegration

- Ziele:
  - Integration verschiedener lokaler Schemata (mit möglicherweise unterschiedlichen Datenmodellen) in ein gemeinsames globales Datenschema
  - Ausräumen aller Inkonsistenzen auf Schemaebene
- Hauptaufgaben:
  - Schema Translation:
    - Abbildung eines Schemas vom Quellschema in äquivalentes Zielschema
    - explizite Abbildung eines jeden Schemas
    - Entwicklung von Abbildungsregeln
  - Schemaanalyse
    - Analyse und Vergleich der Objekte der Schemata
    - Spezifikation von Beziehungen zwischen Objekten der einzelnen Schemata

DWH: ETL

- Im Verlauf der Ladephase müssen folgende Aufgaben bewältigt werden:
  - Zugriff auf die operationalen Datenquellen
  - Erkennen von Änderungen auf den Datenquellen → **E**
  - Transformation der Daten → **T**
  - Bereinigung der Daten → **L**
  - Bilden von Aggregationen
- Probleme beim Ladevorgang:
  - Zugriff auf riesige Datenvolumina bei kleinem Ladefenster
  - sequentielles Laden dauert zu lange
  - Parallelisierung (Füllen der Zielrelationen und Indexerstellung) und inkrementelles Laden notwendig

DWH: ETL - Erkennen von Änderungen

- Erkennen von Änderungen in den Quelldatenbanken
  - kompletter Abzug der Quellen bei jedem Ladevorgang zu aufwendig
  - deshalb: inkrementeller Warehouse Refresh, bei dem nur die seit dem letzten Laden auf den Quellen hinzugekommen Daten geladen werden
  - das Data Warehouse als „eine Folge von Snapshots“
- Unterschiedliche Arten von Informationsquellen:
  - *kooperative Quellen*
    die Änderungen werden von den Quellsystemen explizit dem Warehouse bereitgestellt (z.B. Verwendung von Triggern)
  - *Quellen mit Protokollinformation*
    Identifikation der Änderungen anhand der Logdateien der Quellen
  - *Snapshot-Quellen*
    Extrahierung der Differenz aus zwei Snapshots
  - *anfragbare Quellen*

DWH: ETL - Datentransformation

- “Einfacher” Zugriff auf OLTP- und Altsysteme reicht nicht aus
- Benötigte Funktionalität der Transformation:
  - Änderung des Primärschlüssels
  - Primärschlüsselauflösung bei mehreren Eingabedateien
  - Umformatierung der Daten
  - unterschiedliche Datentypen
  - Bereitstellung von Default-Werten
  - Non-Standard-Formate der Eingabedateien
  - Homonym / Synonym – Problematik
    Homonym: identischer Wert/Begriff für unterschiedliche Semantik
    Synonym: unterschiedliche Werte/Begriffe mit derselben Bedeutung
  - Datenbereinigung

DWH: ETL – Data Cleansing

- Ziele:
  - Datenkonflikterkennung auf Instanzenebene
  - Datenbereinigung auf Instanzenebene
  - Erkennen von Anomalien
  - Erkennen von fehlerhaften Daten und Ausreißern
  - Semantische Probleme: Wortbedeutung;
    Maßeinheit; Genauigkeit, Format, Zeitpunkt
- Ursachen für Dateninkonsistenzen:
  - Fehler im Datenmaterial, die im OLTP-Betrieb nicht zum Tragen kommen
  - Datenkonflikte zwischen Quellsystemen, die redundante Information enthalten (z.B. Kunde "Meier" hat in DB_1 eine andere Telefonnummer als in DB_2)
  - Schemaevolution (z.B. Umstrukturierung des Produktkatalogs, Ausgliederung von Unternehmensteilen)

DWH: ETL – Data Cleansing

- Dateninkonsistenz*erkennung*:
  - Überprüfung semantischer Nebenbedingungen
  - Plausibilitätsprüfungen
  - Überprüfung von Toleranzwerten
  - Überprüfung von Summenwerten
  - Positionsvergleiche
  - nicht alle Datenkonflikte sind erkennbar!
- Dateninkonsistenz*bereinigung*:
  - Anwendungsspezifisches Wissen unbedingt erforderlich!
  - Einsetzen von Nullwerten, Defaultwerten
  - Ermittlung des korrekten Datenwerts
  - Interpolation von fehlenden Datenwerten

DWH: ETL – Data Cleansing

- Elementizing

  Daten: Favoritenstr. 9, A-1040 Wien
  → **Straße**: Favoritenstr.
  → **Hausnr**: 9
  → **PLZ**: 1040
  → **Ort**: Wien
  → **Land**: A - Österreich
- Standardizing

  Standardizierung aller vorliegenden Elemente
- Verifying

  Überprüfung der inhaltlichen Konsistenz. Beispiel: Kombination PLZ + Ort (6010 Wien) ist inkorrekt
- Matching

  Überprüfung Konsistenz durch Vergleich mit anderen Quellen.
  Beispiel: Adresse mit dem Inhalt anderer Datenbanken vergleichen
- (eventuell Householding)

  Überprüfung aller im selben „Haushalt“ wohnenden Personen
- Documenting

  phasenübergreifender Aufbau einer Metadatenbank

DWH / OLAP Schicht

- Analysetoolübersicht

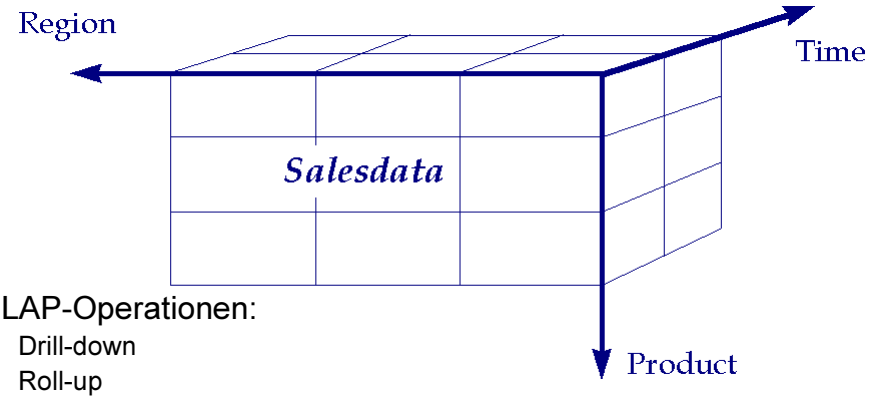
	Statische Abfragen	**Dynamische Abfragen**
Datenumfang gering	Skripten, Anwendungen	OLTP Anfragen
Datenumfang komplex	Reporting	OLAP, Data Mining, Visualisierung

OLAP

- OLAP (on-line analytical processing) als "Anwendungsarchitektur" mit Data Warehouse als Datenquelle
- Business Probleme wie Marktanalysen oder Marktvorhersagen benötigen eine multidimensionale Sicht auf das Datenmaterial (Daten oftmals auch multidimensional modelliert)
- OLAP-Engines erlauben schnellen Zugriff auf Daten durch spezielle Operationen, wie roll-up, drill-down, slice, dice, etc.
- treibende Kräfte:
  - Funktionalität der Tools (Caching, Query Rewriting & Optimierung)
  - Antwortzeiten
  - Security

OLAP – multidimensionale Daten

- OLAP-Cubes:



- OLAP-Operationen:
  - Drill-down
  - Roll-up
  - Slice
  - Dice
  - Pivot / Rotate

OLAP - Guidelines

12 Codd'sche Regeln (Dr. E.F. Codd):

(1) Mehrdimensionale konzeptionelle Sicht
entsprechend den Business Fragestellungen

(2) Transparenz für den Anwender
Entkopplung von heterogenen Quelldaten

(3) Flexible Zugriffsmöglichkeiten
Zugriff nur auf jene Daten, die in der Analyse benötigt werden

(4) Konsistente Geschwindigkeit der Berichtsgenerierung
Datenwachstum oder zusätzliche Dimensionen dürfen die Performance nicht wesentlich verringern

(5) Client-Server Architektur
für maximale Flexibilität, Anpassungsfähigkeit und Interoperabilität

(6) Gleichrangigkeit der Dimensionen
jede Dimension muss dieselben (OLAP-)Operationen erlauben

OLAP - Guidelines

12 Codd'sche Regeln (Dr. E.F. Codd) Fortsetzung:

(7) Dynamische Verwaltung wenig gefüllter Datenwürfel (sparse)
Optimierung des physischen Storage-Modells (ROLAP, MOLAP, HOLAP)

(8) Mehrbenutzerbetrieb

(9) Unbeschränkte, dimensionsübergreifende Funktionen
Gleiche Semantik für OLAP Operationen innerhalb und über Dimensionen übergreifend (z.B. roll-up Summierung)

(10) Intuitive Datenmanipulation
Pivoting, drill-down, roll-up durch „point-and-click"

(11) Flexibles Reporting
(Beispiel: einfache Anpassung der Visualisierung)

(12) Unbegrenzte Anzahl von Dimensionen / Aggregationsebenen

OLAP - Guidelines

Erweiterungen der Codd'schen Regeln:

(13) geeignete DBMS Werkzeuge

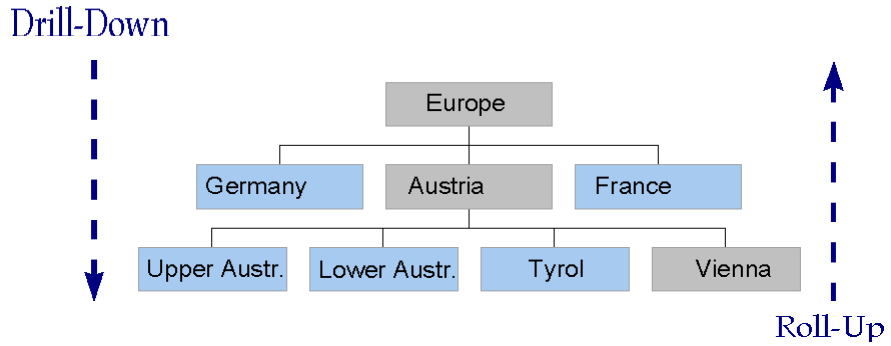
(14) Möglichkeit die Rohdaten zu erhalten
„drill-through"-Funktionalität

(15) Inkrementeller DB Refresh

(16) Strukturiertes SQL Interface

OLAP Operationen

- Roll-Up / Drill-Down



OLAP Operationen

Cube Browser - Sales

	MeasuresLevel					
+ Product Family	Unit Sales	Store Cost	Store Sales	Sales Count	Store Sales Net	Profit
All Products	266.773,00	225.627,23	öS 565.238,13	86837	339.610,90	339.610,90
+ Drink	24.597,00	19.477,23	öS 48.836,21	7978	29.358,98	29.358,98
+ Food	191.940,00	163.270,72	öS 409.035,59	62445	245.764,87	245.764,87
+ Non-Consumable	50.236,00	42.879,28	öS 107.366,33	16414	64.487,05	64.487,05

Double-click a member to drill up or down. Close Help

↓ Drill-down **↑ Roll-Up**

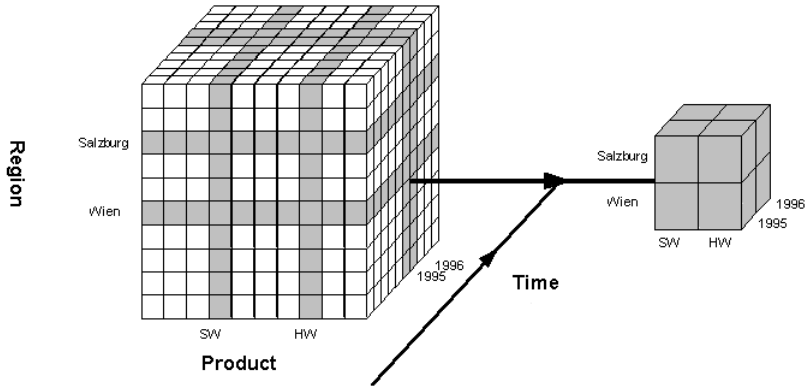
Cube Browser - Sales

		MeasuresLevel				
- Product Family	+ Product Department	Unit Sales	Store Cost	Store Sales	Sales Count	Store Sales Net
All Products	All Products Total	266.773,00	225.627,23	öS 565.238,13	86837	339.610,90
+ Drink	Drink Total	24.597,00	19.477,23	öS 48.836,21	7978	29.358,98
+ Food	Food Total	191.940,00	163.270,72	öS 409.035,59	62445	245.764,87
- Non-Consumable	Non-Consumable Total	50.236,00	42.879,28	öS 107.366,33	16414	64.487,05
	+ Carousel	841,00	595,97	öS 1.500,11	272	904,14
	+ Checkout	1.779,00	1.525,04	öS 3.767,71	569	2.242,67
	+ Health and Hygiene	16.284,00	12.972,99	öS 32.571,86	5310	19.598,87
	+ Household	27.038,00	24.170,73	öS 60.469,89	8862	36.299,16
	+ Periodicals	4.294,00	3.614,55	öS 9.056,76	1401	5.442,21

Double-click a member to drill up or down. Close Help

OLAP Operationen

- Slice / Dice



OLAP Operationen

Cube Browser - Sales

Yearly ... | $150K + | Time | 1997 | Store | USA | Gender | M | Marital ... | All Marital | Custom... | All Custon

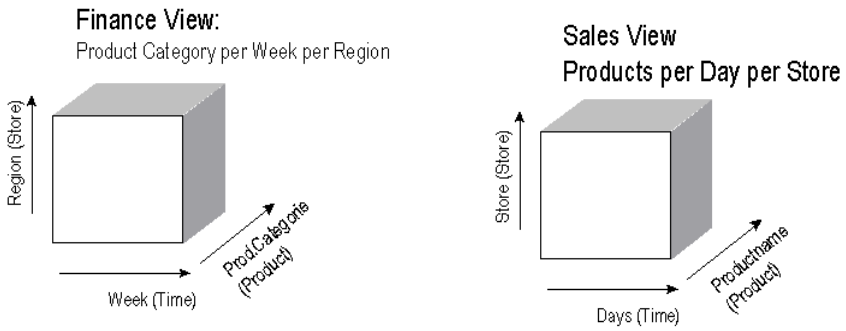
		MeasuresLevel				
- Product Family	+ Product Department	Unit Sales	Store Cost	Store Sales	Sales Count	Store Sales Net
All Products	All Products Total	2.550,00	2.179,47	öS 5.454,79	835	3.275,32
- Drink	Drink Total	250,00	223,12	öS 563,41	85	340,30
	+ Alcoholic Beverages	84,00	82,83	öS 214,70	29	131,87
	+ Beverages	113,00	106,51	öS 256,74	39	150,23
	+ Dairy	53,00	33,77	öS 91,97	17	58,20
+ Food	Food Total	1.830,00	1.541,08	öS 3.859,93	597	2.318,85
+ Non-Consumable	Non-Consumable Total	470,00	415,28	öS 1.031,45	153	616,17

Double-click a member to drill up or down.

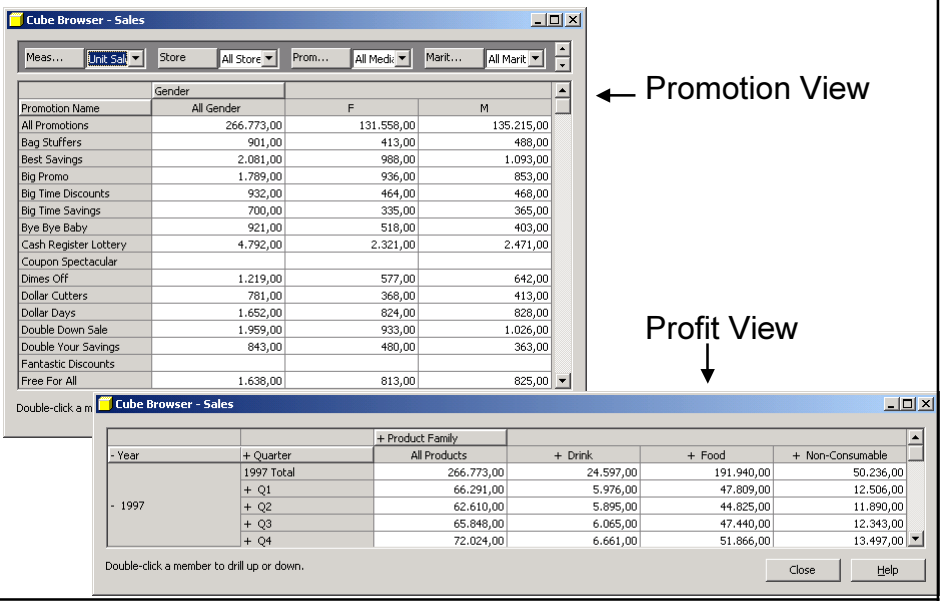
Close Help

OLAP Operationen

- Pivot (Rotate)



OLAP Operationen

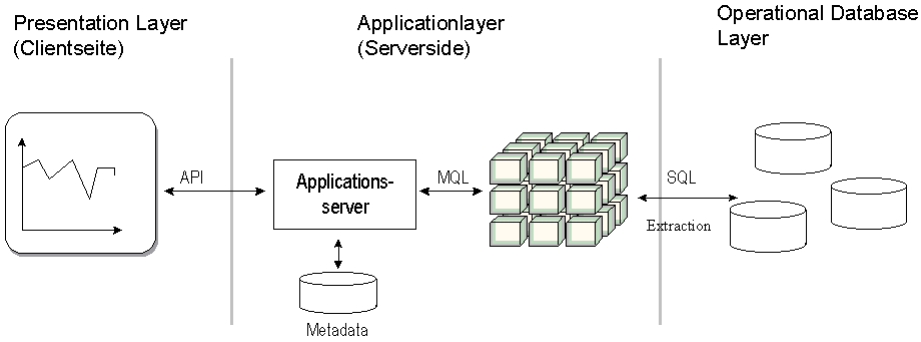


OLAP Anforderungen

- Entwurf einer Datenbank, die „unbekannte“ Abfragen mit sehr guter Performance beantworten kann.
- OLAP Daten (multidimensional) aus dem DWH abgeleitet
- Daten spannen einen multidimensionalen Raum auf
- Eventuell unterschiedliche Speicherungsformen zwischen DWH und OLAP:
  - ROLAP (relationale Datenbank)
  - MOLAP (multidimensionale Datenbank)
  - HOLAP (hybrid OLAP – Kombination ROLAP, MOLAP)
  - DOLAP (Datenbank auf Desktop Computer / client side)

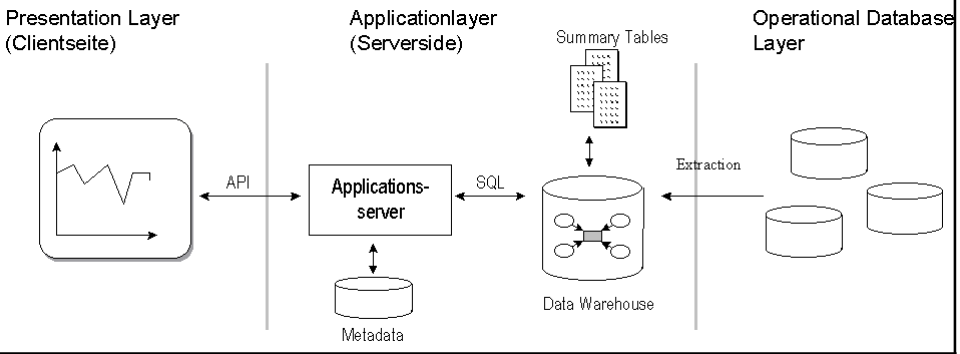
MOLAP - Speicherung

- Multidimensionales OLAP unter Verwendung einer multidimensionalen Datenbank
- Die (Hyper)Cube können direkt aus der Speicherstruktur abgeleitet werden
- Hohe Sparsity, da nicht für jeden Punkt (*n* * *m* * ... * *z* Dimensionen) im multidimensionalen Raum Daten vorhanden sind
- Notwendiger Speicherplatz sehr hoch!



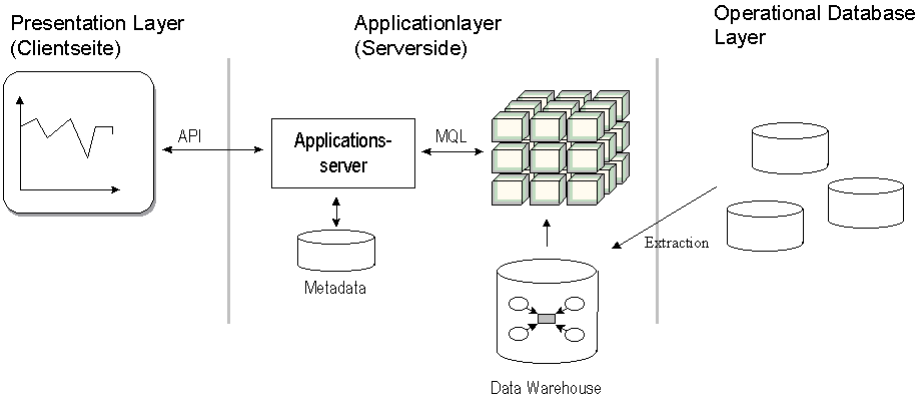
ROLAP - Speicherung

- OLAP basierend auf relationaler Datenbank
- Spezielle Datenmodelle, um multidimensional zu modellieren
  - Bekannteste Datenmodelle: Star Schema, Snowflake Schema
- Optimiert für OLAP Operationen
- Fakt-Tabellen enthalten die relevanten Daten der Geschäftsbereiche
- Dimensions-Tabellen beschreiben die Strukturierung der einzelnen Fakten



HOLAP - Speicherung

- Hybrid OLAP (Kombination ROLAP + MOLAP)
- MOLAP wird aus ROLAP Datenbank (DWH) abgeleitet
- MOLAP für häufig verwendete Daten bzw. dicht befüllte Cube Bereiche
- Bereiche die nur sparse befüllt sind werden nur in ROLAP gespeichert
- Interface ROLAP – MOLAP notwendig



Fakten / Dimensionen

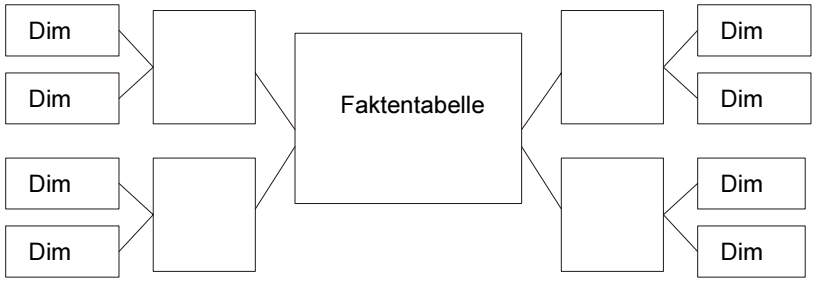
- Fakten:
  - Repräsentieren die primären Geschäftsbereiche
  - Werden nicht geändert sobald sie im DWH sind
  - In bestimmter Granularität gespeichert
- Dimensionen:
  - Referenzieren auf Informationen, die zur Strukturierung / Charakterisierung der Fakten verwendet werden
  - Definieren die Aggregations-Hierarchien
  - Beispiele für Dimensionen: Zeit, Produktgruppen, Regionen

ROLAP: Star Schema

- Denormalisierte Dimensionstabellen
- Zentrale Fakt-Tabelle mit Verweisen in Dimensionstabellen
- Abfragen adressieren die Fakt-Tabelle
- Keine Joins über viele Tabellen notwendig

ROLAP: Snowflake Schema

- Snowflake Schema basiert auf dem Star Schema
- Dimensionstabellen sind normalisiert (3. NF)
- Klarere Strukturierung der Dimensionen (für Leute mit Datenbankerfahrung)
- Verbergung dieser komplexeren Struktur notwendig für die OLAP-Benutzer



Data Mining

- Problematik: Interpretation der Daten und die Generierung von Wissen bei hohem Datenvolumen nicht möglich
  - Datenreduktion oder / und Datenanalyse?
  - Grundidee der Wissensentdeckung: „Wissen aus Daten"
  - nicht nur statistische Annahmen verifizieren kann, sondern „selbständige" Generierung von Hypothesen
  - Data Mining ist keine Verifikation von statistischen Annahmen
  - Goldwäscher Vergleich: mehrmaliger Waschvorgang mit Gitter unterschiedlicher Lochgröße extrahiert „Nuggets" verschiedener Größe
- Begriffsflut: Data Mining, Knowledge Discovery in Databases (KDD), Knowledge Mining, Knowledge Extraction, Datenanalyse (Intelligent Data Analysis)

Data Mining (Definitionen)

« Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques. » (Gartner Group).

« Data mining is the exploration and analysis, by automatic and semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules» (M.J.A. Berry, G. Linoff).

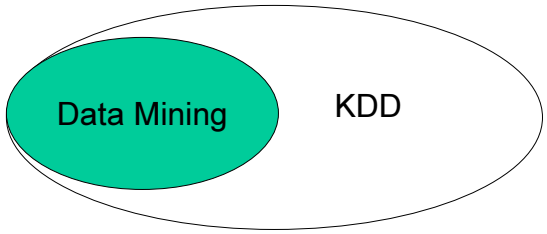
« Data mining the use of advanced statistical tools to reach into a companys existing databases to discover patterns and relationships that can be exploited in a business context» (Trajecta lexicon).

« Data mining is a combine of powerful methods that help reducing costs and risks as well as increasing revenues, by extracting strategical information from the available data. » (T. Fahmy).

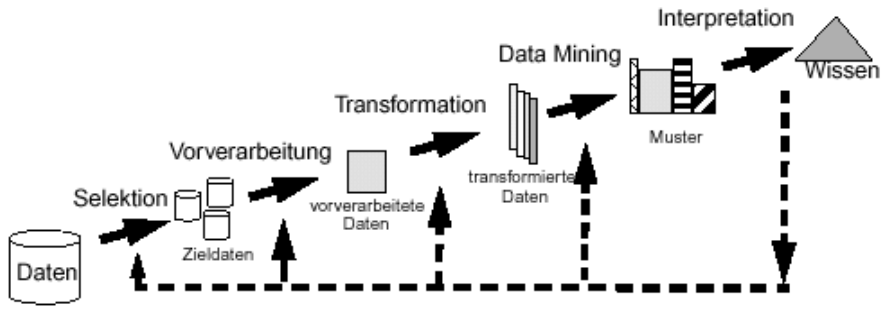
« The automated extraction of predictive information from large databases » (K. Thearling)

Data Mining vs. KDD

- Data Mining ist eine iterative **Suchtechnik**
- Knowledge Discovery in Databases (KDD) ist ein iterativer **Prozeß**, d.h. die Hypothesen des Data Mining Vorgangs werden verifiziert oder interpretiert.



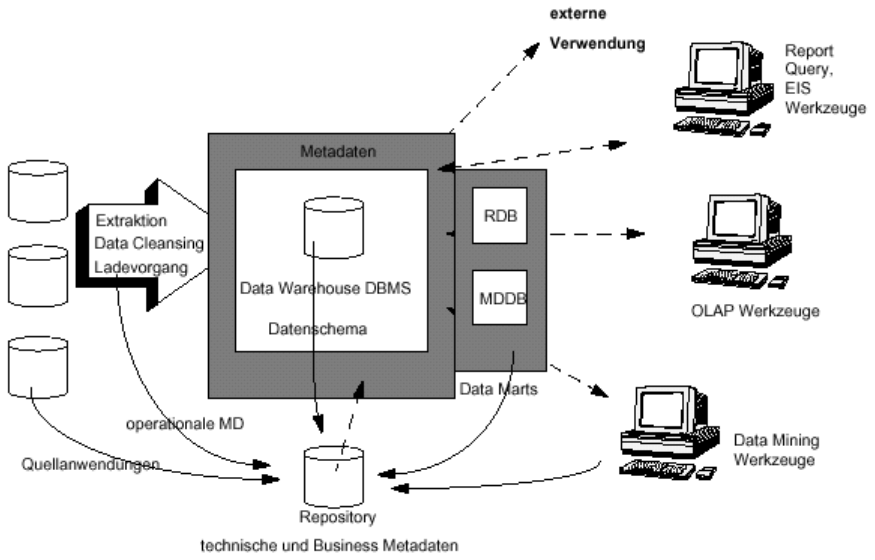
Ablauf des KDD-Prozesses



Metadaten

- Metadaten als Basiskomponente sowohl beim Aufbau als auch beim Betrieb eines Data Warehouses
- Daten zur Beschreibung des DWH: Aufbau, Betrieb und Verwaltung
- Repository als Speichermedium und Zugriffsmedium
  - Beschreibung des gesamten Data Warehouse Systems
  - Namen, Definitionen, Struktur und Inhalt eines DWH und der Endbenutzersichten
  - Identifikation der Datenquellen
  - Integration und Transformationsregeln zum Füllen des DWH
  - Integration und Transformationsregeln für die Endbenutzerregeln
  - operationale Informationen wie Updates, Versionen, Zugriffsverwaltung
  - Verwendung und Performance des DWH
  - Sicherheit / Zugriffsberechtigungen

Metadatenfluß



„Lebenszyklus“ von Metadaten

- **Sammlung:** Identifizierung und Speicherung in einem Repository
  - so automatisch wie möglich
  - Metadaten fallen immer an; zusätzlicher Aufwand ist beim Aufbau des Warehouses am geringsten
- **Verwaltung:** Aktualisierung von Metadaten
  - Metadaten müssen mit der Wirklichkeit übereinstimmen
  - auch so automatisch wie möglich; Probleme vor allem bei Business Metadaten (Abbildung von Business Policies, Rules, etc.)
- **Einsatz:** Metadaten werden dem Benutzer zur Verfügung gestellt
  - benutzerspezifizierte Aussagen und direkte Unterstützung des Benutzers
  - Benutzer von Metadaten: DWH Entwickler; Administratoren; Endbenutzer

Typen von Metadaten

Technische Metadaten - Unterstützung des DWH Betriebs:

- Kontrolle und Management der Komplexität der DWH Umgebung aus einer technischen Perspektive
- Information über das Warehouse für DWH Designer und Administrator
- statische Informationen (Regeln, Permission, Naming Standards, Quellsystemnamen, Orte, Zielnamen wie Tabellen, Spalten, Tablespace, Transformationen, Abbildungsregeln, Definitionen ...)
- dynamische Informationen (Datenextraktion, Cleansing, Ladeskripten, Ladestatistiken, Ladefehler, Zugriffstatistiken, Backupstatistiken, Platzverbrauch im Warehouse, Berichtsverteilung, d.h. Info zum Monitoring und Betrieb des DWH)
- Sicherheitskontrolle im Repository als Datenzugriffsschutz (Zugriffsauthorisierung, Backup, Datenzugriff usw.)

Typen von Metadaten

Business Metadaten

- “Anwender”-Metadata zum einheitlichen Verständnis des Datenmaterials
- zusätzliche Information über Daten (Semantik; vor allem für die Analyse)
- Themengebiete des DWH, Information über Objekttypen, Fragen, Reports, vorgefertigte Anfragen, Komponentenhinweise, operationale Informationen (Datenhistorie, Benutzer, Verwendung)
- Business Rules: Beschreibungen von Beziehungen von Datenelementen
- Betrachtung durch geeignete Metadaten Browser
- Problematik: nicht vorhandene Business Regeln und Unkenntnis, welche Business Rules ein Unternehmen braucht
- kein Automatismus beim Sammeln und Aktualisieren
- Bsp: 5 Prozent aller Vermietungen einer Autovermietung müssen für Tagesvermietungen frei gehalten werden, da Tagesvermietungen teurer sind als Reservierungen => Erkenntnisse für den Analysten für Anfragen an das Datenmaterial

Metadatenmodelle

Metadata Coalition (MDC) / (Object Meta Group) OMG

MDIS (Metadata Interchange Specification, Version 1.1, 1997)

- beschränkt auf die Spezifikation und den Austausch von Datenbankschemata (Transformationslogik auf Metalevel nicht unterstützt)
- keine explizite Unterstützung für Prozesse
- textorientiert
- *monolithisch* aufgebaut, d.h. es gibt keine Trennung zwischen Metamodel und den Instanzen – auch nicht auf physischer Ebene. Sowohl das Model, als auch die Instanzen werden in einem einzigen ASCII-File abgelegt.

Metadatenmodelle

OIM (Open Information Model; Version 1.0; 1999)

- basiert ursprünglich auf einem Industriestandard (unter der Führung von Microsoft und Platinum). Insgesamt über 50 SW-Hersteller involviert (u.a. Informatica, SAS, Brio).
- Version 1.1 befindet sich derzeit in der Reviewphase
- OIM basiert auf UML, XML und SQL und besteht aus fünf anwendungsorientierten Modulen, welche Metadaten beschreiben:
  - Analysis and Design Modul
  - Object and Component Model
  - Business Engineering Model
  - Knowledge Management Model
  - Database and Warehousing Model

Metadatenmodelle

CWM (Common Warehouse Metamodel; Version 1.0; 2000)

- Erweiterung des UML Metamodells und repräsentiert ein objektorientiertes Architektur-Framework für heterogene Systeme.
- Die CWM Spezifikation wurde vor allem unter der Führung von IBM, Oracle, Unisys und Essbase erarbeitet und im Juni 2000 durch die OMG (Object Meta Group) standardisiert.
- CWM definiert in UML die vollständige Spezifikation der Syntax und der Semantik, die notwendig ist, um Metadaten in einer Data Warehouse & Business Intelligence Umgebung auszutauschen.
- Jedes graphische Tool, welches das UML Metamodel versteht, kann zur Modellierung von CWM Modellinstanzen verwendet werden. Die Beschreibungssprache für das CWM Metamodel basiert auf XML – und ist somit ebenfalls standardisiert.

Metadatenmodell – Resümee

- MDIS: sehr rudimentär, hauptsächlich Data Mapping
- OIM: de-facto Industriestandard
- CWM
  - Mächtigster Standard
  - Schnittstellen zum Datenaustausch sehr genau spezifiziert
  - CWM Enablement Showcase (UML Forum, Tokyo, 03.2001): 6 verschiedene Hersteller haben gezeigt, dass der Austausch von Metadaten basierend auf CWM zwischen zehn verschiedenen Data Warehouse Tools und Repositories möglich ist.
- Zusammenschluss von MDC und OMG lässt mittelfristig eine Vereinigung von OIM und CWM erwarten

DWH Administration

- Sicherheitsmanagement
- Aktualisierungsmonitoring aus unterschiedlichsten Quellen
- Datenqualitätsüberprüfung
- Management und Aktualisierung von Metadaten
- Archivierung der DWH Verwendung
- Replizierung und Verteilung von Daten
- Backup und Recovery

DWH Projekt

- 60 - 80 % der DWH-Projekte schlagen „fehl“
  - Vergleich: Chaos Report (Standish Group): 26% der SW-Projekte innerhalb Zeit und Ressourcen fertig. Basis: ca. 30000 untersuchte SW-Projekte in den USA in verschiedensten Anwendungsfeldern
- kein Patentrezept möglich, da zu viele Einflussfaktoren
- Typische Wasserfallmethode nicht anwendbar:
  - Anforderungen nicht stabil und nicht vollständig im Vorfeld zu evaluieren
  - keine vollständige Spezifikation vorhanden
  - Übergreifendes Verständnis fehlt
  - Programmierung nicht die Hauptproblematik

DWH Projektrollen

- Executive Sponsor
- Projektmanager
- Technischer Leiter
- DWH Architekt
- Datenanalyst
- DWH Tool Spezialist
- Anwendungsentwickler
- Datenqualitätsanalyst
- Business Projekt Leiter
- Business Analyst
- Subject Area Specialist
- Reportentwickler
- Operationsspezialist

Allgemeine Aufgaben/Rollen:
Datentransformation, Datenextraktion, Netzwerkanbindung, Dokumentation, Systemadministrator, Datenbankadministrator, Sicherheitsexperte

DWH Projektphasen

(1) Definition Zielrichtung
- Triebfeder: Business, Strategie, Taktik und Benutzer
- Teamzusammenstellung

(2) Definition technische Architektur
- Komponenten und Werkzeuge festlegen

(3) Aufbau der DWH-Infrastruktur
- Design, Aufbau und Füllen eines Metadaten Repository
- Kaufen, Installieren, Testen der Hardware and Software

(4) Erste Iteration
- Ansatz zum Testen (Prototyp)

(5) Aufsetzen des DWH-Verwaltungsprozesses
- Extraktion, Laden, Metadaten usw.

(6) Ausbau der Analyseumgebung

(7) Vervollständigen der Iterationen (→ Punkt 4)
- Testen, Tuning, Training

10 Fehler beim Aufbau eines DWH

Quelle: DW Institute (1995)

(1) Start mit der falschen Finanzierungskette
- Geldgeber als Projektleiter

(2) zu hoch angesetzte Erwartungen

Data Warehouse Projekt besteht immer aus 2 Phasen:
- Verkaufsphase
- Umsetzung

(3) firmenpolitische Fehler
- keine Versprechungen wie “DWH für bessere Entscheidungen”

(4) zu viele Daten laden (nur weil sie vorhanden sind)
- zu viele unnötige Daten führen zu großem DB und HW Aufwand, hohen Kosten

(5) Glaube, dass DWH Design dem OLTP Design entspricht
- Ziele sind unterschiedlich; a) inhaltlich und b) verarbeitungstechnisch!

10 Fehler beim Aufbau eines DWH

Fortsetzung:

(6) Verwendung eines Projektleiters der technik-orientiert ist (anstatt einen benutzer-orientierten)
- Data Warehouse ist eine Dienstleistung

(7) Fokussierung auf interne Datenquellen (ohne Beachtung der externen)
- viele Informationen eines Unternehmens kommen von außen

(8) Überlieferung der Daten mit überlappenden und konfusen Definitionen
- Konsensfindung als Basisaktivität beim Aufbau eines DWH

(9) Glaube an Performance, Kapazitäts- und Erweiterungsfähigkeit der Systeme
- Hardware, Software und Netzwerk

(10) Nach dem Aufbau des Data Warehouses gibt es keine Probleme mehr
- neue Daten, neue Tools → DWH ist eine “Reise” und kein Ziel

10 Fehler beim Aufbau eines DWH

Fortsetzung:

(11) ... es ist ein Fehler an 10 Fehlern festzuhalten

- Fokussierung auf OLAP, Data Mining und Reporting
- Alert/ Exceptionreporting Systeme als neue Generation der Informationsgewinnung

Definition nach Inmon

A Data Warehouse is a ...

- subject-oriented,
- integrated,
- time-variant,
- nonvolatile

... collection of data in support of management's decision making process.

(W. H. Inmon)

Fragen ?