Data Warehousing

Matthias Conrad
97I1

DB Oberseminar

Inhaltsverzeichnis

- Einführung
- DWH - Architektur
- Akquisition
- Multidimensionales Datenmodell
- Materialisierte Sichten
- Metadaten
- Ausblick
- Referenzen

Einführung DWH

Einführung

„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)

Einführung

- Unternehmen stehen heute immer mehr Daten elektronisch zur Verfügung

  → Großer Druck diese zur Unterstützung des Entscheidungsprozesses einzusetzen

  → Viele Daten können nicht genutzt werden (da nicht in interpretierbarer Art vorliegend, sich diese auf verschiedenen Systemen befinden)

- Braucht Systeme die Daten zu Informationen aufwerten

→ Technologien, die Daten aus verschiedenen Quellen in fundierte, einheitlich strukturierte Informationen auswerten, werden unter dem Begriff des ***Data Warehousing*** zusammengefasst.

Einführung

- Achtziger Jahren kamen erste Datenmodelierungsmethoden auf

  → Erlaubte die Anforderungen an die Daten und die dazu benötigten Strukturen formal zu dokumentieren

- Notwendigkeit einer Struktur und Architektur in Bezug auf die Datenbeschaffung → um Übersicht zu behalten

- Ende der achtziger Jahre Unterscheidung zwischen operativen und analytischen Informationssystemen

  → *OLTP* (On-line Transaction Processing)
  → *OLAP* (On-line Analytic Processing)

Einführung

- operative – analytische Datenbanken

	Operative DB	**Analytische DB**
Einsatz	operatives Geschäft	Analyse, Entscheidungsunterst.
Daten	aktuell, isoliert, detailliert	historisiert, integriert, detailliert, aggregiert
Verarbeitung	Transaktionen, häufige Änderungen	Anfragen, Ergänzungen
Eigenschaften	Konsistenz, Vollständigkeit	Qualität, Richtigkeit
Grösse	MB...GB	GB...TB

Einführung

- Analytische DB hauptsächlich entworfen um die Ausführung von komplexen meist read-only Anfragen zu unterstützen

  → Anfragedurchsatz und Antwortzeit, wichtiger als Transaktionsdurchsatz

- Die Industrie war nun vor neue Anforderungen gestellt

  → Technologische Einschränkungen, vor allem um Informationen von verschiedenen heterogenen Systemen zusammenzubringen, behinderten die Entwicklung von OLAP-Systemen.

Einführung

- Die Data Warehousing Technologie zielt darauf hin, Lösungen für diese Probleme zu liefern

- Seit Mitte der neunziger Jahre ist DWH ein fester Bestandteil unserer Informationsgesellschaft

Einführung

- Die Anforderung an ein Data Warehouse sind :

  → Informationen einer Unternehmung zugänglich machen
  → Der Inhalt der Daten muss klar, verständlich sein
  → Schnelle Ausführung von Anfragen mit minimaler Wartezeit
  → Informationen konsistent halten
  → Sämtliche Informationen sind vollständig und erklärt
  → Anpassbar und flexibel
  → Das Informationsvermögen sollte gut vor Missbrauch geschützt sein
  → Es soll die beste Grundlage für den Entscheidungsprozess bieten, so dass sich die getroffenen Entscheidungen auch bewähren

Einführung

- Anwender: Manager
  Abteilungsleiter
  Fachkräfte

- Formen der Bereitstellung :

  → Query-Ansätze: frei definierbare Anfragen und Berichte
  → Reporting: Zugriff auf vordefinierte Berichte
  → Redaktionell aufbereitete, personalisierte Informationen

Einführung

- Betriebswirtschaftliche Anwendungen

Analyse

→ Detaillierte Analyse der Daten zur Untersuchung von Abweichungen oder Auffälligkeiten

Planung

→ Unterstützung durch explorative Datenanalyse
→ Aggregierung von Einzelplänen

Einführung

Kampagnenmanagement

→Unterstützung strategischer Kampagnen
→Kundenanalyse, Risikoanalyse

Beispiel: Wal Mart (www.wal-mart.com)

- Größe: ca. 25 TB
- Täglich bis zu 20.000 DWH-Anfragen
- Basis für Warenkorbanalyse, Kundenklassifizierung

Einführung

- Wissenschaftliche Anwendungen

  → Projekt Earth Observing System (Klima- und Umweltforschung)
  täglich ca. 1,9 TB meteorologischer Daten
  Aufbereitung und Analyse (Data Mining)

  → Öffentlicher Bereich: DW mit Umwelt- oder geographischen Daten (z.B. Bodenanalysen)

DWH – Architektur

DWH – Architektur

- Anforderungen an ein Data Warehouse

  → Unabhängigkeit zwischen Datenquellen und Analysesystemen

  → Dauerhafte Bereitstellung integrierter Daten

  → Mehrfachverwendbarkeit von Daten

  → Durchführung von Auswertungen

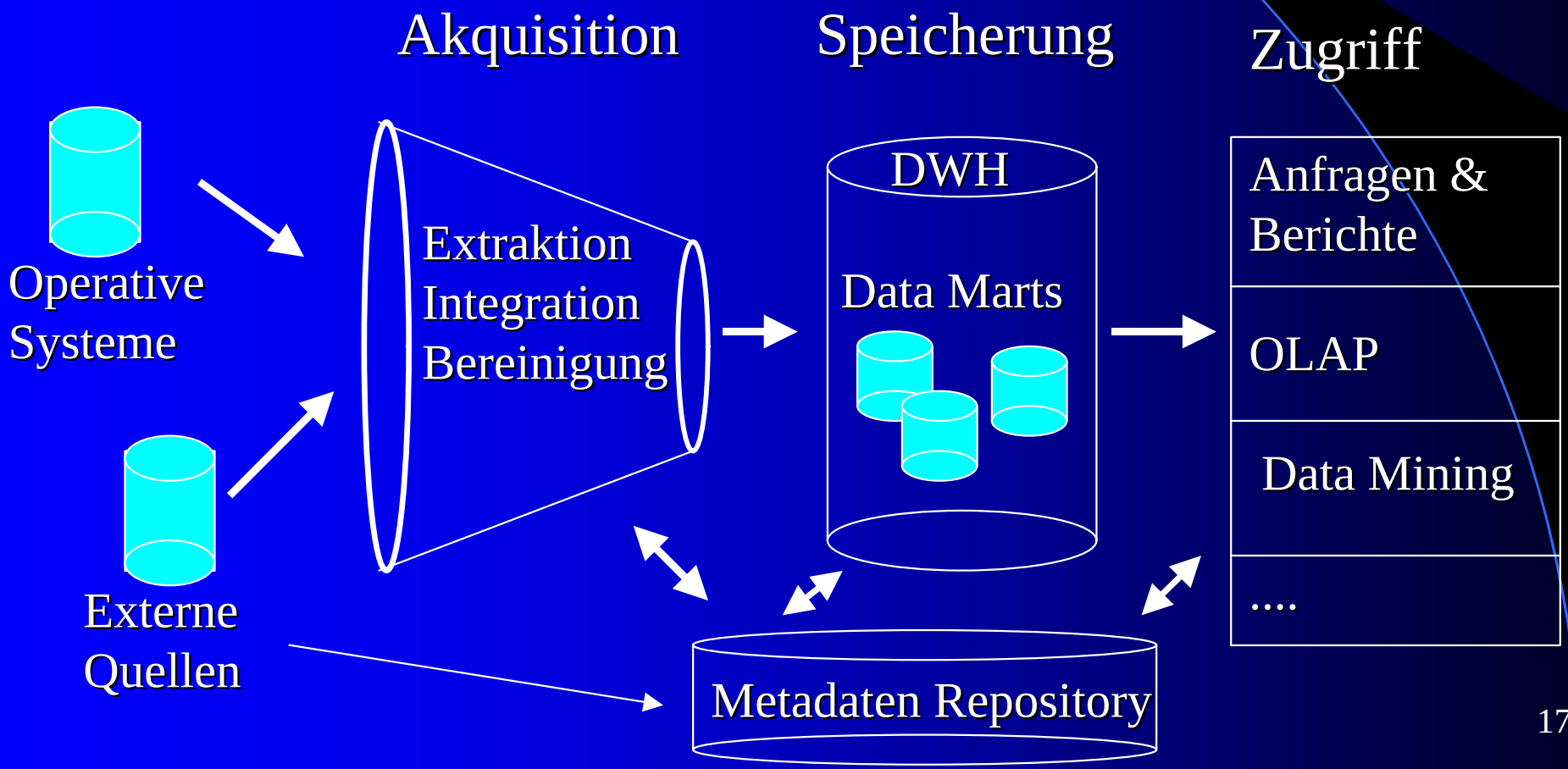
  → Erweiterbarkeit des DWH

  → Zweckorientiert

  → Unterstützung individueller Sichten

DWH – Architektur

Ein DWH System beinhaltet das Data Warehouse sowie alle Komponenten, die für die Entwicklung, den Unterhalt und den Zugriff auf das DWH benötigt werden.



DWH – Architektur

- Daten werden von operativen Systemen / externen Quellen bezogen
- Die Quellen sind autonome Komponenten
- DWH System hat keine Kontrolle über Inhalt und Form der Daten in den Quellen
- Der Bereich Akquisition dient der Datenaufbereitung

  →Dies beinhaltet : bereinigen
  vervollständigen
  kombinieren
  aggregieren

- Daten werden im Anschluss in das eigentliche DWH überführt

DWH – Architektur

- Im eigentlichen DWH werden die Daten abgespeichert
- Meistens in Data Marts (Daten Märkten)

  → Data Marts sind Teilmengen eines DWH
  → Natürliche Trennung (Unternehmensbereiche ...)

- Durch die Erstellung von Data Marts können die Abfragen beschleunigt werden

- In vielen Fällen werden Data Marts eingesetzt weil :

  → Anwenderzugriffswerkzeuge
  → Zugriffsrestriktionen
  → besondere Datenschutzbestimmungen es erfordern

DWH – Architektur

- Zugriffskomponente aller Applikationen
  - → OLAP
  - → Data Mining
- Der Endbenutzer des Data Warehouse greift nur über diese Zugriffskomponente auf die Daten im System zu

→ Hier werden also die Informationen gewonnen

- Weitere Informationen über OLAP / Data Mining siehe
  - → 11.01.01 Maik Kurzhals (OLAP)
  - (→ 17.01.01 Katja Wachsmuth (Data Mining))

DWH – Architektur

- Zusätzlich gibt es → *Metadaten – Managementkomponente*
- Definiert, pflegt und arbeitet mit den verschiedenen Typen von Metadaten
- Allgemein sind Metadaten als Daten über Daten definiert
- DWH gibt es verschiedene Typen von Metadaten
  - → Informationen über Struktur
  - → Informationen über Semantik der Daten
  - → Informationen über den Unterhalt und den Zugriff auf das DWH
- Dient dem schnellen und sicheren Auffinden der benötigten Daten / Informationen

Akquisition

Akquisition

- Ziel: Beschaffung und Aufbereitung von Daten für das DWH
- Aufgaben: Extraktion
  Transformation
  Bereinigung
  Vervollständigung
  Integration
- Datenakquisition sehr aufwendig 60 % bis 80 % der Projektzeit
  - → Unterschiedliche Art der Datenquellen
  - → Qualität der extrahierten Daten
  - → Datenvolumina
  - → Limitiertes Zeitfenster

Akquisition

- Datenextraktion

  - Aus operativen Quellen ausgelesen → in Staging Area zwischengespeichert

  - Meist über Standard-Middleware ausgelesen (z.B. ODBC)

  - Bei älteren Datenquellen muss ein spezieller Extraktor erstellt werden

  → Aktualisierung

- Datentransformation

  - Daten liegen unterschiedlichen, quellenabhängigen Formaten vor

  - Umwandlung in einheitliche Struktur

Akquisition

- Datenbereinigung

  - Ungenügende Qualität der Daten (unvereinbare Datenformate, fehlende Werte, unlesbare Teile, Duplikate, Tippfehler)

    „garbage in, garbage out“

  - Zu beachten : Korrektheit
    Konsistenz
    Vollständigkeit
    Aktualität
    Glaubwürdigkeit
    Redundanz
    Verständlichkeit
    Verfügbarkeit

Akquisition

- Datenvervollständigung
  - Behandlung fehlender Werte
  - Berechnen abgeleiteter Werte
  - Bilden von Aggregationen
- Datenintegration
  - Zusammenmischen der Daten anhand der definierten Beziehungen
  - Datenintegration über Mapping-Tabellen

Akquisition

- Laden ins DWH
  - → Letzter Schritt der Datenakquisition
  - → Laden der Daten aus Staging Area ins DWH
  - → Logisches Schema des DWH wird beeinflusst
  - → Daten werden mittels Regeln ins DWH eingefügt
    (dabei indexiert und sortiert)
- Monitoring
  - → Nach der Initialisierung des DWH fortlaufende Überwachung der Quellsysteme

Multidimensionales Datenmodell

Multidimensionales Datenmodell

- Datenmodell ausgerichtet auf Unterstützung der Analyse

  → Entscheidungsprozeß

- Betrachtung der Kennzahlen aus unterschiedlichen Perspektiven (zeitlich, produktbezogen) → Dimensionen

- Kennzahlen/Fakten (engl. facts):

  → Numerische Messgrößen
  → Beschreiben betriebswirtschaftliche Sachverhalte
  → Beispiele: Umsatz, Gewinn, Verlust

  Typen: Additive Fakten
  Semi-additive Fakten
  Nicht-additive Fakten

Multidimensionales Datenmodell

- Dimension:
  - → Beschreibt mögliche Sicht auf die Kennzahl / Fakten
  - → Endliche Menge von n (n >= 2) Dimensionselementen die eine semantische Beziehung aufweisen
  - → Beispiele: Produkt, Geographie, Zeit
- Hierarchien in Dimensionen
  - → Einfache Hierarchien
  - → Parallele Hierarchien
  - → Würfel

Multidimensionales Datenmodell

- Würfel (engl. cube)

  → Grundlage der multidimensionalen Analyse

  → Kanten (Dimensionen)

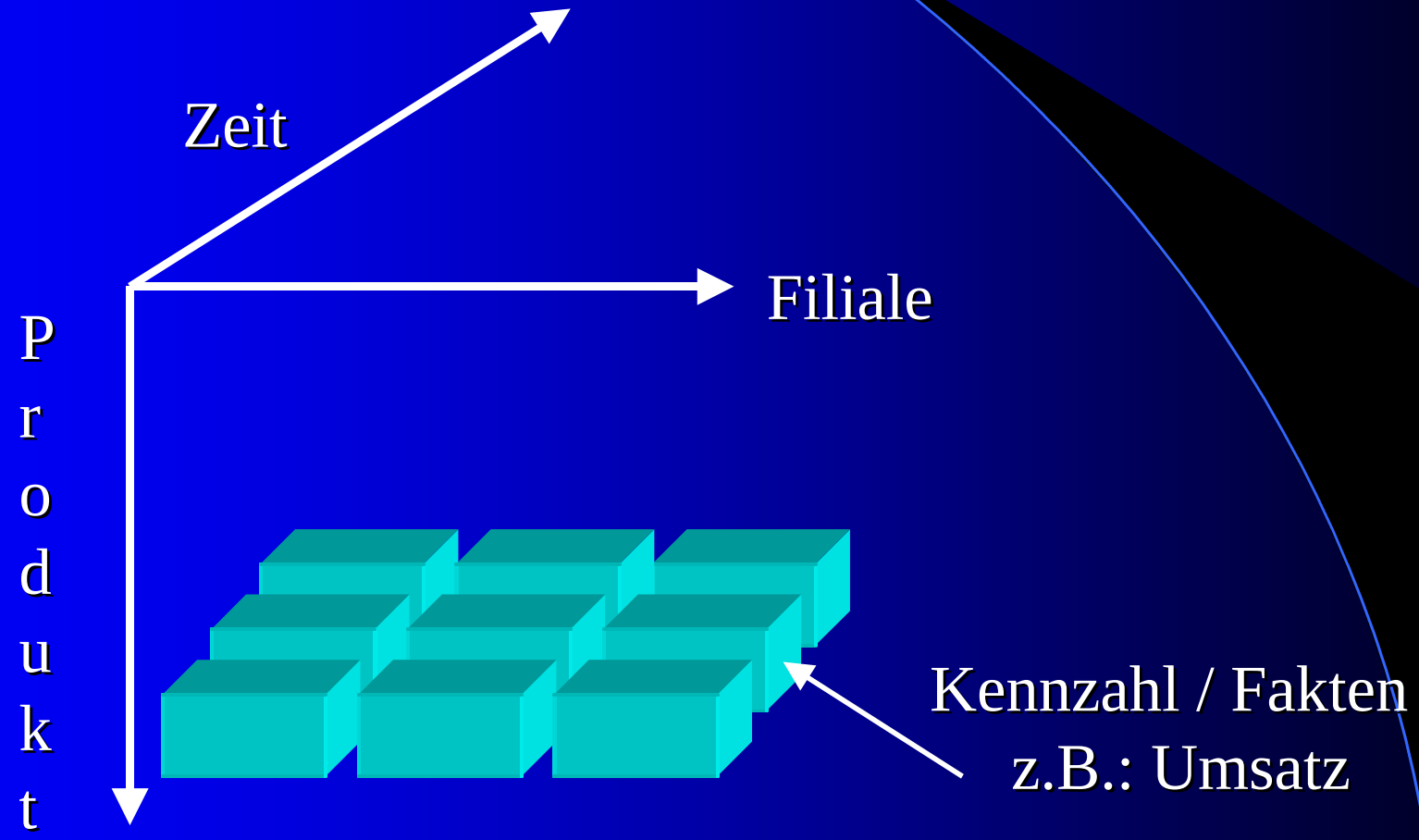
  → Zellen (ein oder mehrere Kennzahlen)

  → Visualisierung:

  2 Dimensionen: Tabelle
  3 Dimensionen: Würfel
  >3 Dimensionen: Multidimensionale Domänenstruktur

Multidimensionales Datenmodell

- Würfel am Beispiel eines Kaufhauses



Multidimensionales Datenmodell

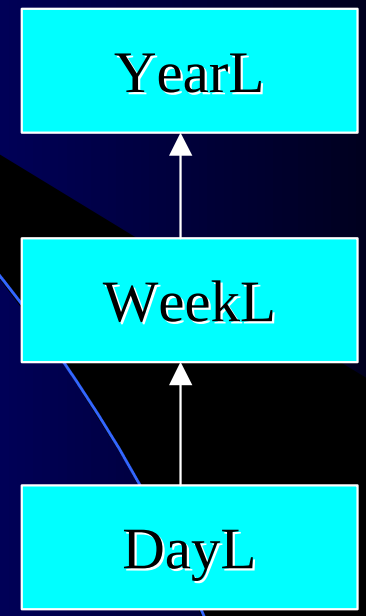
Days Dimension

```
Create Table Days (
        date_desc       DATE,
        day_WH          NUMBER ,
        day_of_year     NUMBER ,
        day_of_month    NUMBER ,
        month_desc      VARCHAR2(9),
        month_number    NUMBER ,
        quarter         NUMBER ,
        week_of_year    NUMBER ,
        year            NUMBER )
TABLESPACE USERS;

ALTER TABLE Days ADD
CONSTRAINT Dayl_UK UNIQUE(day_WH);
```

Multidimensionales Datenmodell

```
CREATE FORCE DIMENSION Days_DIM
        LEVEL YearL IS DAYS.year
        LEVEL QuarterL  IS DAYS.quarter
        LEVEL MonthL    IS DAYS.day_of_month
        LEVEL WeekL     IS DAYS.week_of_year
        LEVEL DayL      IS DAYS.day_WH
  HIERARCHY DMQY (
        DayL CHILD OF
        MonthL CHILD OF
        QuarterL CHILD OF YearL )
  HIERARCHY dwy (
        DayL CHILD OF
        WeekL CHILD OF YearL )
ATTRIBUTE MonthL DETERMINES (ml_month_desc,ml_month_number)
ATTRIBUTE DayL DETERMINES (dl_date_desc,dl_day_of_year);
```



Multidimensionales Datenmodell

- OLAP-Operationen zielen auf die multidimensionalen Datenstrukturen ab

→ Standardoperationen

Pivotierung / Rotation
Roll-Up, Drill-Down
Drill-Across
Slice, Dice

Materialisierte Sichten

Materialisierte Sichten

- Ein mögliche Realisierung eines DWH ist die Speicherung von Daten aus einer oder mehreren operativen Datenbanken in Form von materialisierten Sichten

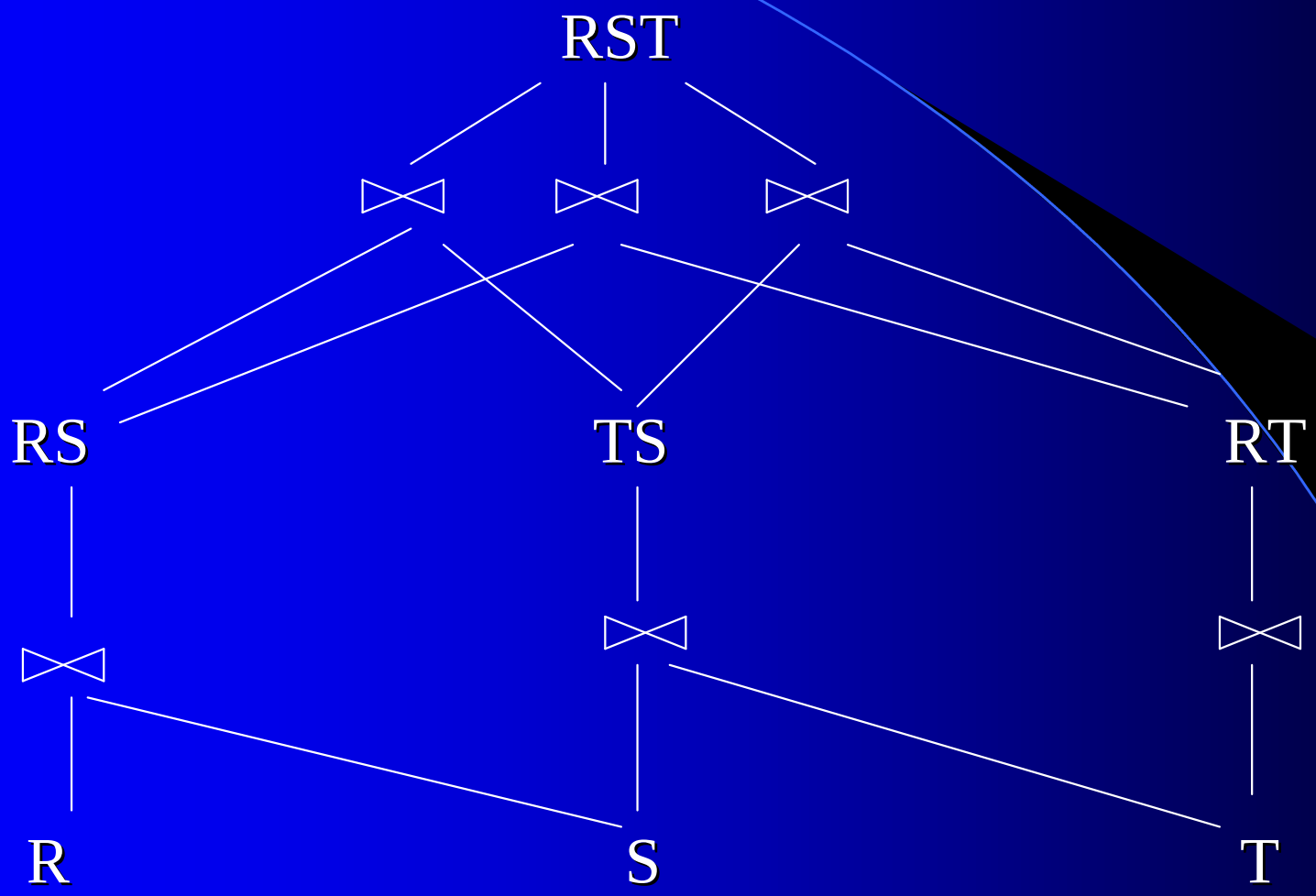
  → Um so schnellen Zugriff auf die Daten zu ermöglichen unabhängig von der Verfügbarkeit der Datenquellen

  → Sonst Konsistenz schnell verloren / Daten unbrauchbar werden

- Eine Sicht wird aus einer Funktion und einer oder mehreren Basisrelationen generiert und ist selbst wieder eine Relation

- Sicht wird physisch auf dem Datenträger gespeichert (Unabhängigkeit gegenüber den Quelldaten)

Materialisierte Sichten



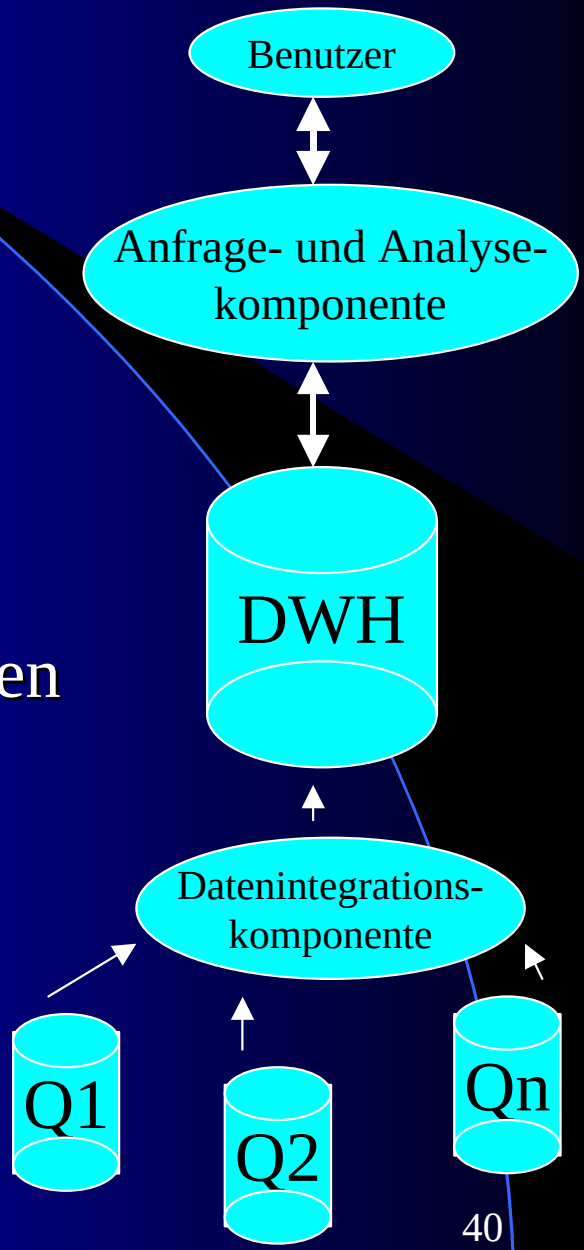
Materialisierte Sichten

CREATE VIEW <Sichtname> [(Spalte1, Spalte2...)] AS
SELECT <Tabellenname Spaltennamen>
FROM <Tabellenname/ Viewname>
WHERE <Suchbedingung>

CREATE VIEW *spiel_verk_2000* **AS**
SELECT *f.Stadt, f.Manager, v.Verkaufs_id, v.Monat,*
g.Gegenst_id, g.Gegenst_name, l.Linien_id, l.Verkaufs_Preis
FROM *Filiale f, Verkauf v, Linie l, Gegenstand g*
WHERE *f.Filialen_id = v.Filialen_id* **and**
v.Verkaufs_id = l.Verkaufs_id **and**
l.Gegenst_id = g.Gegenst_id **and**
f.Land = „D“ **and**
v.Jahr = „1999“ **and**
g.Kategorie = „Spielwaren“

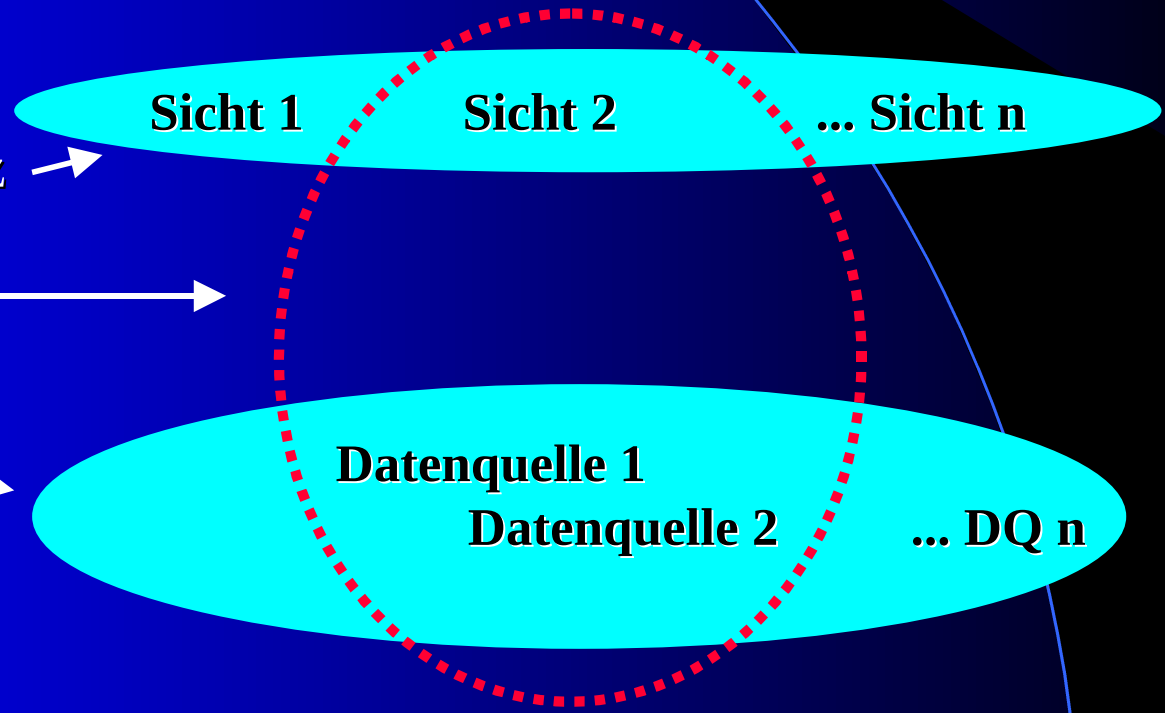
Materialisierte Sichten

- Erstellung/Initialisierung von Sichten
  - Nach Modellierung der MS müssen diese im System spezifiziert werden
  - Integrationskomponente beauftragt, die entsprechenden Daten aus den Quellen zu holen
  - Diese nach den Vorgaben die MS zu erstellen
  - Diese werden im DWH abgespeichert
  - Sind nun für den Benutzer verfügbar



Materialisierte Sichten

- Konsistenz
  - Sicherstellung der Konsistenz nach Änderungen in den Quellbeständen
  - Drei verschiedene Konsistenzbereiche :



- Multiple Sichtenkonsistenz
- Sichtenkonsistenz
- Quellenkonsistenz

Materialisierte Sichten

- Oftmals Änderungen im Quelldatenbestand Auswirkung auf das DWH
- Die MS müssen so angepasst werden, damit Konsistenz erhalten bleibt
- Methoden :
  - → DWH sendet Anfragen an die Quelle, worauf diese die Antworten auf die Anfragen ans DWH liefert
  - → Die Quelle sendet automatisch Updates an das DWH, woraufhin das Update auf die Sichten angewendet wird

Metadaten

Metadaten

- Metadaten -- „Daten über den Daten,“ nehmen im DWH eine Schlüsselrolle ein

  → Beschreiben z.B. die im DWH vorhandenen 0 und 1 und ermöglichen deren Interpretation.

- Zu den Metadaten zählen :

  - Konzeptuelle Entwurf
  - Programmcodes
  - Sicherheitseinrichtungen
  - Namen und Eigenschaften von Tabellen
  - Weitere Informationen über Tabellen
  - Aus welchen operationalen Systemen die Daten ins DWH gelangt sind

Metadaten

- Klassifikation von Metadaten

→ Nach Benutzung

Technische Metadaten	Geschäftliche Metadaten
Logisches und konzeptuelles Datenmodell	Namen der Tabellen und Attribute (in klaren Bezeichnungen)
DWH -Tabellennamen, - Schlüssel und - Indizes (z.B. in der Form von Code)	Zeitpunkt der Aktualisierungen
Verantwortlichkeiten und Zugriffsrechte innerhalb des DWH	Abfragungs- und Navigationspfade

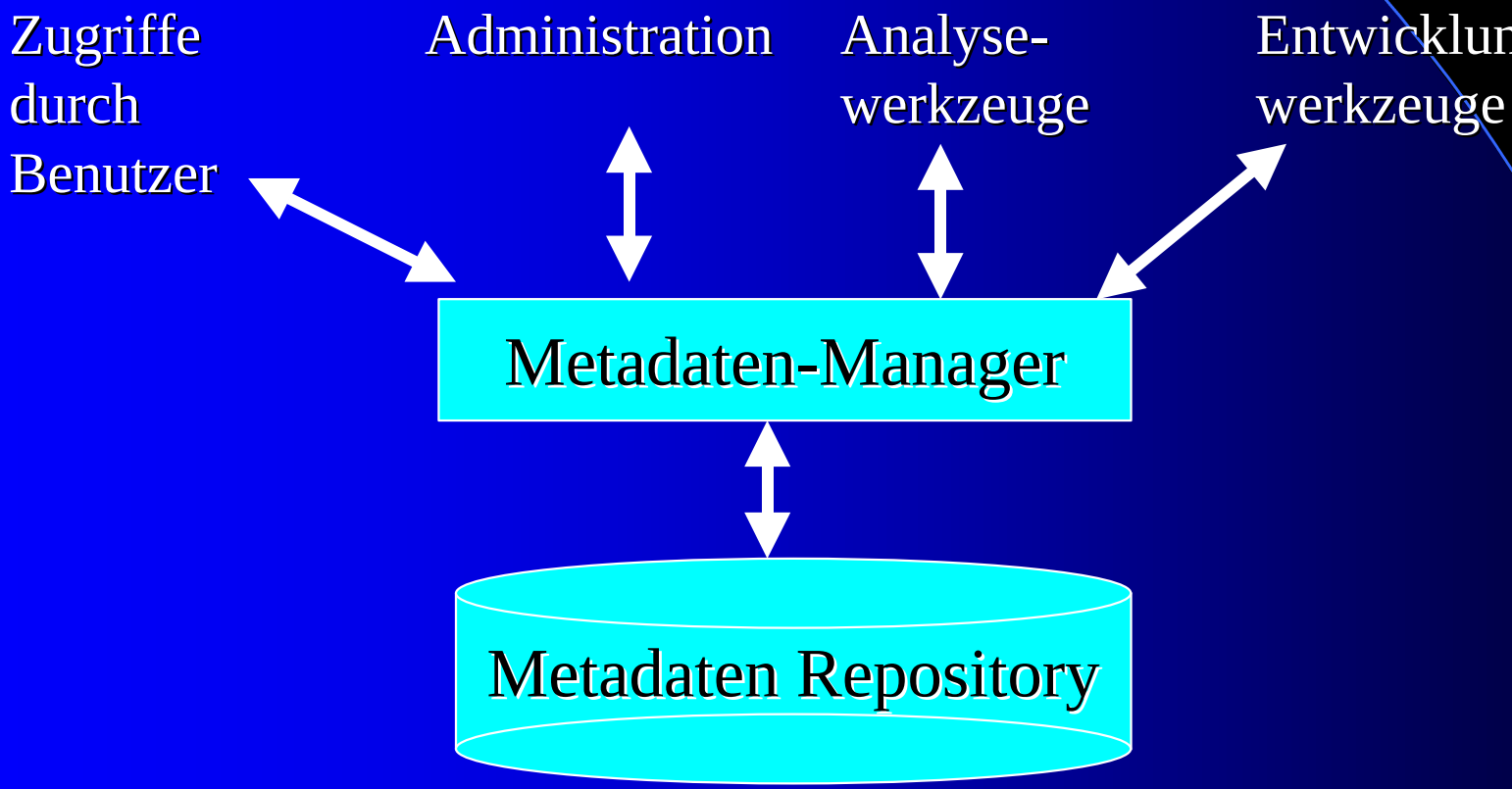
Metadaten

- Verfeinerung der Klassifikation
  - Typ : → Primärdaten
    → Metadaten für Prozesse
  - Abstraktion : Modellierung konzeptuell, logisch und physisch
  - Erstellungs- / Verwendungszeitpunkt

Erstellungszeitpunkt	**Verwendungszeitpunkt**
Entwurfsmetadaten	CASE-Werkzeuge
Aufbaumetadaten	Transformations- und Qualitätsregeln
Benutzungsmetadaten	OLAP

Metadaten

- Die Metadaten werden im Metadaten-Repository gespeichert
- Dem Repository vorgelagert befindet sich der Metadaten-Manager über den der gesamte Zugriff und Austausch erfolgt



Metadaten

- Das Metadaten Repository soll allen Benutzergruppen einen optimalen Informationsgewinn aus dem DWH ermöglichen

- Diverse Standards :
  - → Open Information Model (OIM)
  - → Common Warehouse Model (CWM)

welche auf unterschiedliche Art und Weise Metadaten strukturieren und austauschen

Ausblick

Ausblick

- DWH raschen Veränderungen unterworfen
- Der Markt wächst extrem schnell
- Ende 1998 betrug er 5 Milliarden US$

  → 2002 auf 21 Mrd. US$ prognostiziert
- DWH sehr dynamisch
- Kontinuierliche Weiterentwickelung
- Unternehmen ohne die Informationen aus einem DWH müssen Wettbewerbsnachteile in Kauf nehmen

Referenzen

- Wolfgang Martin „ Data Warehousing “
- TECHNET.ORACLE.COM
- ORACLE.COM
- OWB Users Guide
- Institut für Informatik der Universität Zürich
- W.H. Inmon „ Building the Warehouse “
- Bauer, Günzel „ Data Warehouse – Architektur, Entwicklung, Anwendung “