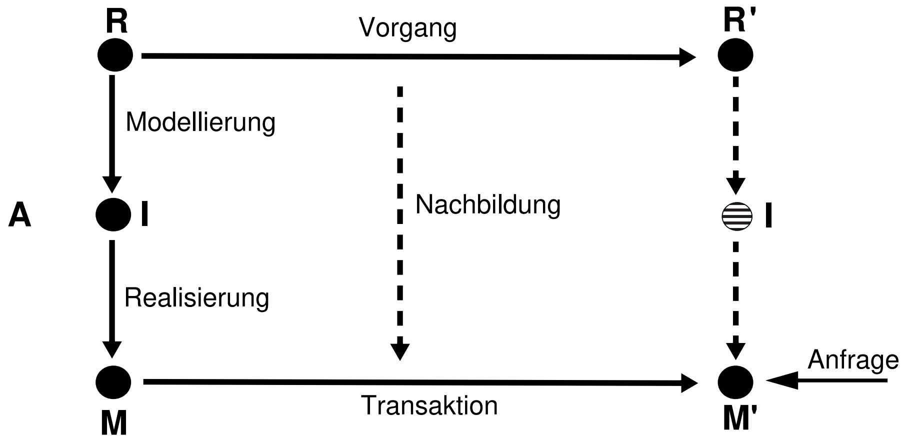
1. Einordnung von DBS

- **Miniwelt[1] – modellhafte Abbildung**
- **Information – Was ist das?**
- **Was ist ein Informationssystem?**
  - Komponenten eines rechnergestützten Informationssystems
  - Anforderungen an ein betriebliches Informationssystem
- **Datenbanksysteme – erste Annäherung**
  - Aufgaben und Eigenschaften
  - Beispiel: Relationenmodell (strukturierte Daten)
- **Unstrukturierte und semi-strukturierte Daten**
  - Information-Retrieval-Systeme
  - HTML, XML
  - Suchqualität
- **Klassen von DB-Anwendungen**
  - Was ist Transaktionsverarbeitung ?
  - Nutzung von Data Warehouses
  - Einsatzszenarien von DBS

1. Ein Datenbanksystem verwaltet Daten einer realen oder gedanklichen Anwendungswelt. Diese Daten gehen aus Informationen hervor, die stets aus den Sachverhalten und Vorgängen dieser Anwendungswelt durch gedankliche Abstraktionen (Abbilder, Modelle) gewonnen werden. Sie beziehen sich nur auf solche Aspekte des betrachteten Weltausschnitts, die für den Zweck der Anwendung relevant sind. Ein solcher Weltausschnitt wird auch als *Miniwelt* (Diskurswelt) bezeichnet.

Miniwelt – modellhafte Abbildung

- **Grobe Zusammenhänge**



R:	Realitätsausschnitt (Miniwelt)
I:	Informationsmodell (zur Analyse und Dokumentation der Miniwelt)
M:	DB-Modell der Miniwelt (beschrieben durch Objekt- und Beziehungsmengen sowie Integritätsbedingungen usw.)
A:	Abbildung aller relevanten Objekte und Beziehungen ➥ Abstraktionsvorgang

- **Transaktion:**
  - bildet Vorgang in **R** im DBS nach und
  - garantiert ununterbrechbaren Übergang von **M** nach **M'**
    ➥ implementiert durch Folge von DB-Operationen
  - DB-Anfragen beziehen sich auf **M** bzw. **M'**
- **Integritätsbedingungen:**
  - Zusicherungen über **A**, **I** und **M**: A_1: R→I, A_2: I→M
    ➥ Ziel: möglichst gute Übereinstimmung von **R** und **M**
  - **Idealfall:** Die DB ist zu jeder Zeit ein Abbild (Modell) der gegebenen Miniwelt

Miniwelt – modellhafte Abbildung (2)

- **Transaktionskonzept**
  - führt ein neues Verarbeitungsparadigma ein
  - ist Voraussetzung für die Abwicklung betrieblicher Anwendungen (*mission-critical applications*)
  - erlaubt „Vertragsrecht" in rechnergestützten IS zu implementieren

- **Welche Eigenschaften von Transaktionen sind zu garantieren? (ACID-Paradigma)**
  - **Atomicity (Atomarität)**
    - TA ist kleinste, nicht mehr weiter zerlegbare Einheit
    - Entweder werden alle Änderungen der TA festgeschrieben oder gar keine („alles-oder-nichts"-Prinzip)
  - **Consistency**
    - TA hinterläßt einen konsistenten DB-Zustand, sonst wird sie komplett (siehe Atomarität) zurückgesetzt
    - Zwischenzustände während der TA-Bearbeitung dürfen inkonsistent sein
    - Endzustand muß die Integritätsbedingungen des DB-Modells erfüllen
  - **Isolation**
    - Nebenläufig (parallel, gleichzeitig) ausgeführte TA dürfen sich nicht gegenseitig beeinflussen
    - Alle anderen parallel ausgeführten TA bzw. deren Effekte dürfen nicht sichtbar sein
  - **Durability (Dauerhaftigkeit)**
    - Wirkung einer erfolgreich abgeschlossenen TA bleibt dauerhaft in der DB erhalten
    - TA-Verwaltung muß sicherstellen, daß dies auch nach einem Systemfehler (HW- oder System-SW) gewährleistet ist
    - Wirkungen einer erfolgreich abgeschlossenen TA kann nur durch eine sog. kompensierende TA aufgehoben werden

Information – Was ist das?

- **Beobachtung**

  Die Praxis der Information und Kommunikation entwickelte sich rasant, ohne daß der Informationsbegriff von einer Theorie hinreichend geklärt wurde. Es gibt keine Theorie oder gar Philosophie der Information.

  ➥ Eine resignierende Schlußfolgerung[1]:
  Die Definition des Begriffes „Information" ist nicht möglich.
  Jeder Versuch dazu setzt ähnliche Begriffe voraus, beispielsweise „Wissen" oder „Kommunikation". Diese Definition wäre damit <u>zyklisch</u>.

- **Erklärungsversuche (philosophisch, technisch, pragmatisch)**

  1. „Information ist neben Materie und Energie etwas Drittes"

  2. Information für Menschen über seine Umwelt:
     Information setzt den Menschen über seine Außenwelt in Kenntnis, ist also der „Stoff", der Erkenntnis ermöglicht

  3. Informationstheorie nach Shannon:
     Statistischer Informationsbegriff (Entropie einer Nachrichtenquelle):

     $$H(p_1,\ldots,p_n) = -\sum_{i=1}^{n} p_i \cdot \log p_i$$

  4. Information und Nachricht
     Die übermittelte Nachricht ist dann von Bedeutung, wenn wir eine Abbildung kennen, die sie mittels einer Interpretationsvorschrift α auf eine Information abbildet:

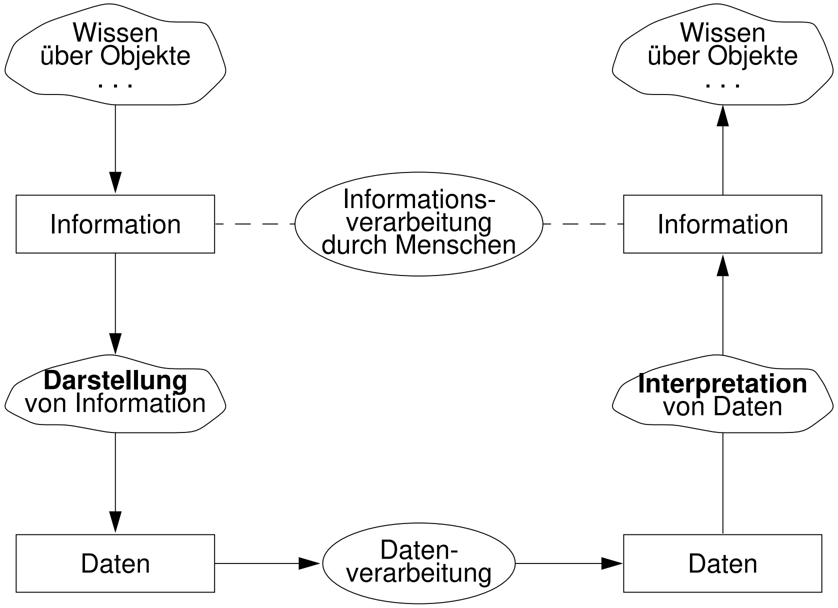
     $$\mathbf{N} \xrightarrow{\alpha} \mathbf{I}$$

1. Bauer, F.L., Goos, G.: Informatik – Eine einführende Übersicht, 3. Auflage, Springer-Verlag, 1982

Information – Was ist das? (2)

- **Erklärungsversuche**

5. Informationsbegriff nach DIN: **Erklärungsmodell**
   (auch für Nachrichtenaustausch zwischen Sender und Empfänger)



Information: subjektive Welt der bewerteten Daten

Daten: objektive Welt der nicht-interpretierten Daten

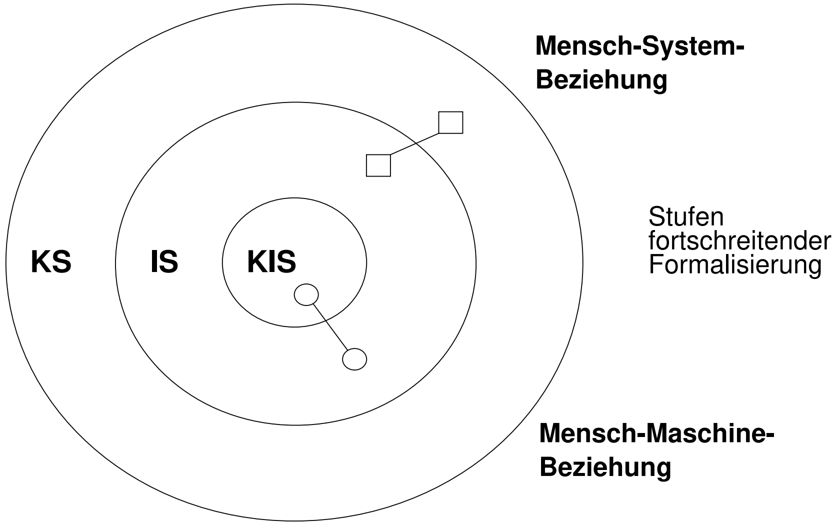
6. Pragmatische Festlegung in der BWL[1]
   Information: **Angaben über Sachverhalte und Vorgänge** (Hansen)
   - „Dabei kann man Information im Sinne von als zweckgerichtetes Wissen zur Vorbereitung und Durchführung von Handlungen verstehen“.
   - „Eine andere Sichtweise ergibt sich aus der Betrachtung der Information als Produktionsfaktor“.

1. Wirtschaftsinformatik-Lexikon, Gabler-Verlag, 1997

Was ist ein Informationssystem?

- **Charakterisierung eines IS nach Senko:**

“The purpose of an information system is to provide a *relatively exact, efficient, unambigous* model of the significant resources of a real world enterprise.”



Formalisiertes Kommunikationssystem = Informationssystem

- **(Vage) Definitionen:**

Ein Informationssystem[1] (IS) besteht aus Menschen und Maschinen, die Informationen erzeugen und/oder benutzen und die durch Kommunikationsbeziehungen miteinander verbunden sind.

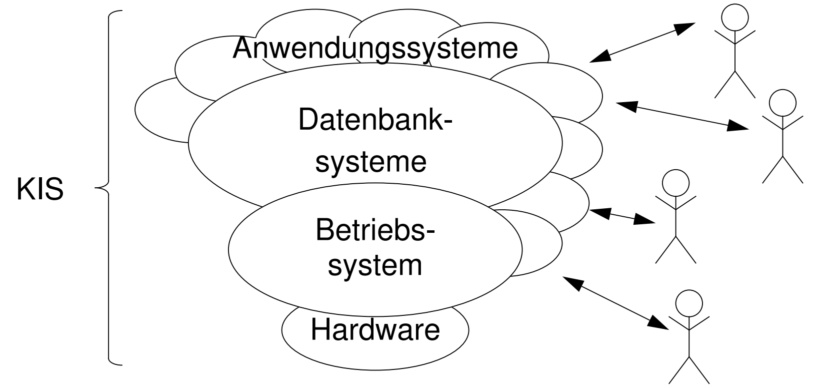
Ein betriebliches IS dient zur Abbildung der Leistungsprozesse und Austauschbeziehungen im Betrieb und zwischen dem Betrieb und seiner Umwelt.

Ein rechnergestütztes IS ist ein System, bei dem die Erfassung, Speicherung und/oder Transformation von Informationen durch den Einsatz von EDV teilweise automatisiert ist. In der betrieblichen Praxis besteht es typischerweise aus einer Menge unabhängiger Systeme, die zusammen die angestrebte Leistung erbringen
(*KIS: kooperatives Informationssystem*)

1. Als „System im weiteren Sinne“ gilt (a) eine Menge von Elementen (Systembestandteilen), die (b) durch bestimmte Ordnungsbeziehungen miteinander verbunden und (c) durch klar definierte Grenzen von ihrer Umwelt geschieden sind; von „Systemen im engeren Sinne“ oder „technischem System“ spricht man, wenn sowohl die Außenwirkungen des Systems insgesamt wie auch seine Binnenstruktur (d. h. die Ordnungsbeziehungen der Systembestandteile) durch Zielfunktionen bestimmt sind (H. Wedekind).

Rechnergestützte Informationssysteme

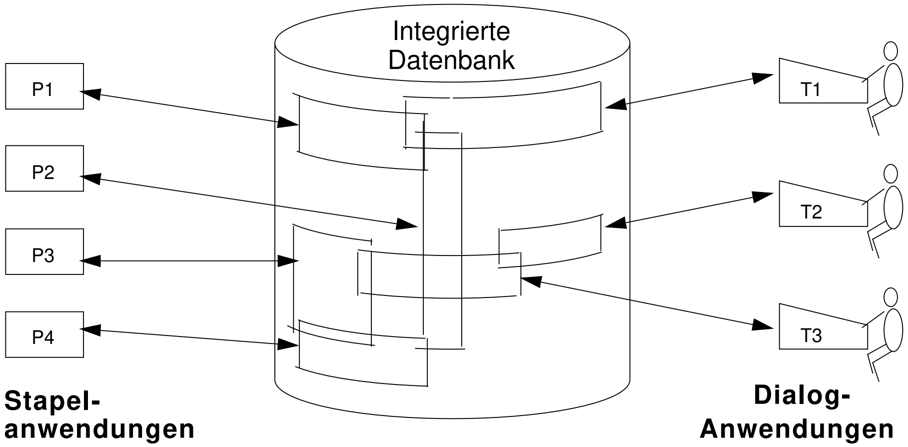
Datenbanksysteme (DBS): (zentrale) Hilfsmittel für KIS



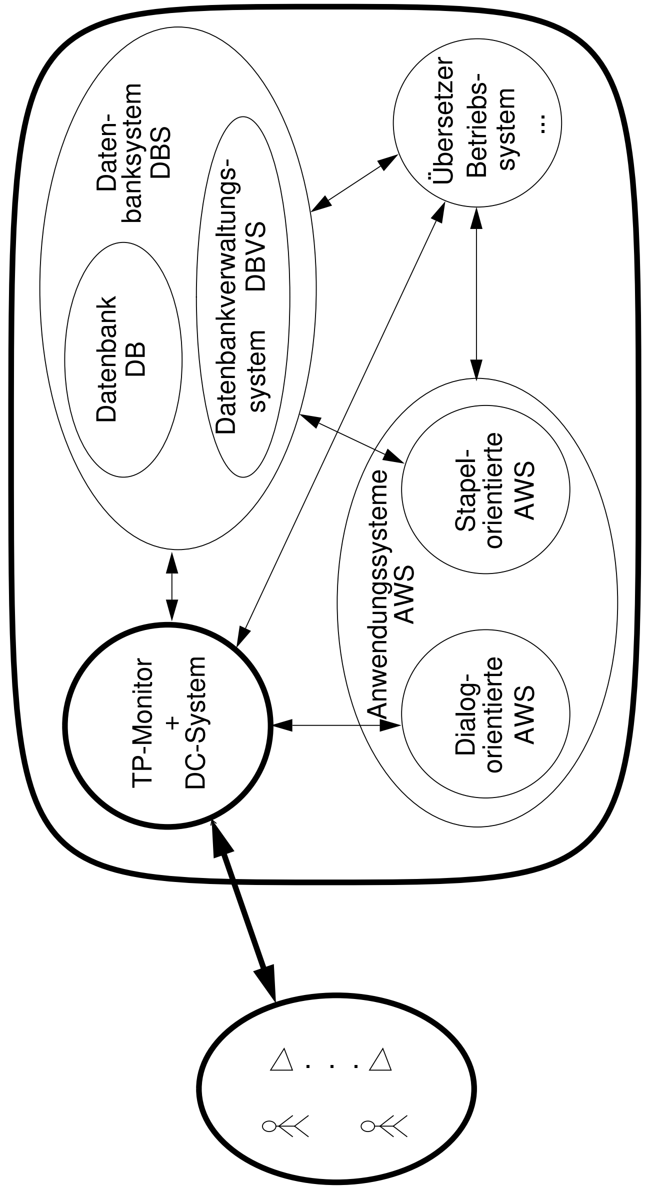
DBS = DB + Datenbankverwaltungssystem (DBVS, DBMS)

Eine ***Datenbank*** ist eine Sammlung gespeicherter operationaler Daten, die von den Anwendungssystemen eines bestimmten Unternehmens benötigt werden.

Ein ***DBVS*** ist ein standardisiertes Softwaresystem zur Definition, Verwaltung, Verarbeitung und Auswertung der DB-Daten. Es kann mittels geeigneter Parametrisierung an die speziellen Anwendungsbedürfnisse angepaßt werden.



Erweiterung der Sicht eines KIS

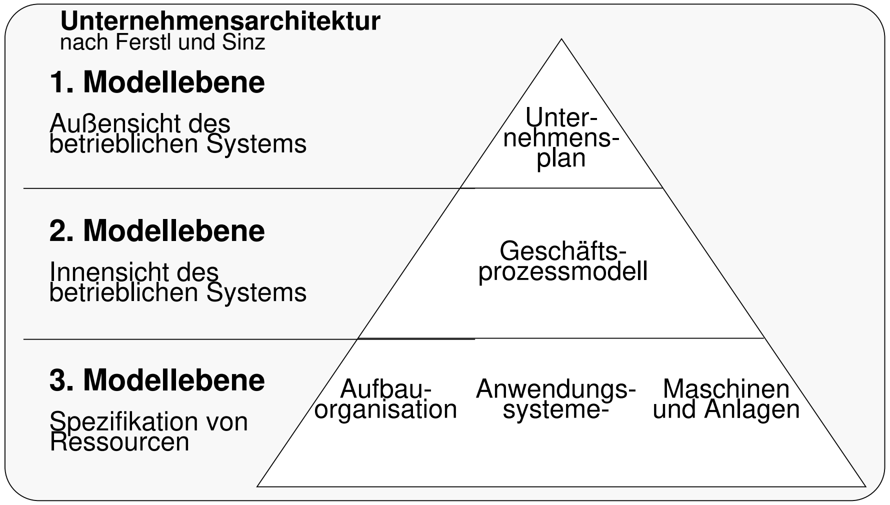


Dialogorientierte AWS: Transaktionsanwendungen mit Interaktion des Endbenutzers, harte Zeitrestriktionen, Terminal-E/A für Anforderungen und Ergebnisse

Stapelorientierte AWS: (Transaktions-) Anwendungen ohne Interaktion des Endbenutzers, keine Zeitrestriktionen, Datei-E/A für Anforderungen und Ergebnisse

Aufgaben betrieblicher Informationssysteme

- **Betriebliche Informationssysteme**
  - zentraler und strategisch bedeutsamer Bestandteil von Unternehmen
  - systematischer Aufbau und Ausrichtung an den Unternehmenszielen
  - Orientierung an Unternehmensarchitektur als dreischichtiger Pyramide



- **Administrative und operative Ebene als Fundament**
  - wird gebildet von Mitarbeitern (eingegliedert in Organisationseinheiten und Stellen einer Aufbauorganisation), Anwendungssystemen sowie Maschinen und Anlagen
  - Interaktion zwischen ihnen dient der Verfolgung definierter Ziele
- **Planungs- und Kontrollebene**
  - Interaktionen der 3. Ebene werden in Form von Geschäftsprozeßmodellen formuliert
  - Erreichen der Geschäftsziele wird überwacht
- **Strategische Ebene**

  Unternehmensplan formuliert Ziele und weitere Randbedingungen, die durch Ausführung von Geschäftsprozessen erreicht bzw. eingehalten werden sollen

Aufgaben betrieblicher Informationssysteme (2)

- **Unterscheidung nach Aufgaben des**
  - betrieblichen Lenkungssystems (Planung, Steuerung und Kontrolle)
  - betrieblichen Leistungssystems (Administration, Disposition und Durchführung)
- **Aufgaben und Typen (Beispiel)**
  - Administrationssysteme dienen der Rationalisierung und (Teil-)Automatisierung vorhandener Abläufe
  - Dispositionssysteme sollen die kurzfristige betriebliche Entscheidungsfindung vereinfachen/übernehmen
  - Planungssysteme unterstützen die mittel- bis langfristige Entscheidungsfindung (Erzeugung alternativer Pläne, weitreichndere Auswirkung)
  - Kontrollsysteme dienen dem Erkennen von außergewöhnlichen und daher bemerkenswerten Situationen (Datenkonstellationen). Sie erhalten von Administrationssystemen Ist-Daten, um Ist-Soll-Abweichungen erkennen zu können

    ➥ hier: Erarbeitung der technischen Grundlagen
- **Was sind typische Aufgabenbereiche?**

Typ/Aufgabe	Produktion	Beschaffung Lagerhaltung	Vertrieb	Personal
Administration	Betriebsdaten-erfassung	Lagerver-waltung	Kundenver-waltung	Personal-verwaltung
Disposition		Bestell-disposition		
Kontrolle	Fertigungs-leitstand		Touren-planung	
Planung	Absatz-/Kapa-zitätsplanung		Marketing-planung	Personalein-satzplanung

Zielsetzungen für betriebliche IS

- **Anforderungen an ein Informationssystem in einem Unternehmen sind unterschiedlich, je nachdem, ob Aufgaben**
  - der operierenden Ebene (Sachbearbeitung)
  - der planenden Ebene (mittleres Management)
  - der strategischen Ebene (Unternehmensleitung)

  zu lösen sind.

- **Verbesserung aller Prozesse und Aufgaben der operierenden Ebenen** des Unternehmens durch Auskunfts-, Berichts-, Buchungs-, Produktions-, Steuerungs-, Vertriebs- und Anwendungssysteme

  Kennzeichen: *Verarbeitung großer Datenmengen und große Änderungshäufigkeit der Daten*

- **Unterstützung und evtl. Teilautomatisierung aller Prozesse und Aufgaben der planenden Ebene** durch:
  - benutzerorientierte Bereitstellung von Informationen
  - Suche und Auswertung von Daten im Dialog
  - Automatisierung von Routine-Entscheidungen
  - Einsatz von mathematisch-statistischen Methoden

  Kennzeichen: teilweise unvorhersehbarer Informationsbedarf, verdichtete Daten, kein Änderungsdienst

- **Unterstützung der strategischen Ebene** durch Bereitstellung von Daten für einen **überwiegend nicht vorhersehbaren Informationsbedarf.**

Beispiele für Informationssysteme

Universitätsdatenbank

Die Universitätsdatenbank ist die Sammlung aller für die Abwicklung der an einer Universität anfallenden Verwaltungsaufgaben benötigten Daten.

Eine Universität gliedert sich i. allg. in mehrere Fachbereiche, denen sowohl die Studenten als auch die Professoren zugeordnet sind.

Die Studenten belegen verschiedene Vorlesungen von Professoren und legen bei ihnen Prüfungen ab.

Typische Anwendungen sind z. B.:
Immatrikulation der Studienanfänger, Rückmeldung der Studenten, Ausfertigen von Studentenausweisen und Studienbescheinigungen, Stundenplanerstellung und Planung der Raumbelegung, Ausstellen von (Vor)diplomzeugnissen, Exmatrikulationen, Statistiken über Hörerzahlen, Raumauslastung, Prüfungsergebnisse, etc.

Datenbank eines Produktionsbetriebes

In einem Produktionsbetrieb werden Daten über die verschiedenen Abteilungen und deren Beschäftigte mit ihren Familienangehörigen gespeichert.

Die Angestellten arbeiten an verschiedenen Projekten mit. Jedes Projekt benötigt für seine Durchführung bestimmte Teile. Jedes Teil kann von Lieferanten bezogen werden. Die Projekte werden jeweils von einem Projektmanager geleitet.

Die in einem Betrieb hergestellten Endprodukte setzen sich i. allg. aus mehreren Baugruppen und Einzelteilen zusammen.

Typische Anwendungen sind z. B.:
Einstellung und Entlassung von Personal, Lohn- und Gehaltsabrechnung, Bestellung und Lieferung von Einzelteilen, Verkauf von Fertigprodukten, Lagerhaltung, Bedarfsplanung, Stücklistenauflösung, Projektplanung.

Beispiele für Informationssysteme (2)

Datenbank einer Fluggesellschaft

Eine Fluggesellschaft fliegt verschiedene Flughäfen an. Auf diesen Flugstrecken werden Flugzeuge bestimmter Typen mit dafür ausgebildetem Personal eingesetzt. Die Piloten haben Flugscheine jeweils nur für einige wenige Flugzeugtypen. Außer den Piloten gibt es noch anderes Bord- sowie Bodenpersonal.

Die Flugbuchungen der Passagiere sowie das Anfertigen der Passagierlisten werden ebenfalls automatisiert durchgeführt.

Typische Anwendungen sind z. B.:
Flugbuchungen von Passagieren, Personaleinsatzplanung, Materialeinsatzplanung, Flugplanerstellung, Überwachung der Wartelisten, Gehaltsabrechnung.

Datenbank einer Bank

Eine Bank gliedert sich gewöhnlich in mehrere Zweigstellen auf. Die Angestellten der Bank gehören jeweils fest zu einer bestimmten Zweigstelle. Auch die Bankkunden sind immer einer Zweigstelle zugeordnet. Es sind Daten über die verschiedenartigen Konten der Bankkunden bereitzustellen, wie z. B. Girokonten, Sparkonten, Hypothekenkonten, Kleinkreditkonten, Wertpapierkonten, etc.

Typische Anwendungen sind z. B.:
Buchung von Zahlungsvorgängen auf den verschiedensten Konten, Einrichten und Auflösen von Konten, Kreditgewährung bzw. Bereitstellen von Daten über die Kreditwürdigkeit eines Kunden, Zinsberechnung und -verbuchung, sowie alle Vorgänge der Personalverwaltung wie z. B. Gehaltsabrechnung.

Zur Rolle rechnergestützter Informationssysteme im Bankenbereich:

“In banking, by contrast, the data actually is the inventory – the two are synonymous. In increasingly many cases, the DB transaction is the financial transaction. There are no real, tangible tokens (greenbacks) moved as a result of the monetary transfer transaction. If the data is bad, money is lost or created. There is no possibility of counting the money (bits) in order to verify the status. Fiscal responsability dictates that creating or destroying money – even temporarily – is unacceptable.”
(Mike Burman, Bank of America)

Datenbanksysteme – erste Annäherung

- **Allgemeine Aufgaben/Eigenschaften von DBS**
  - Verwaltung von persistenten Daten (lange Lebensdauer)
  - effizienter Zugriff (Suche und Aktualisierung) auf große Mengen von Daten (GBytes – TBytes)
  - flexibler Mehrbenutzerbetrieb
  - Verknüpfung / Verwaltung von Objekten verschiedenen Typs (➥ typübergreifende Operationen)

- **Datenstrukturen**
  - <u>formatierte</u> Datenstrukturen, feste Satzstruktur
  - Beschreibung der Objekte durch Satztyp, Attribute und Attributwerte (S_i/A_j/AW_k)
  - jeder Attributwert AW_k wird durch Beschreibungsinformation (Metadaten) A_j und S_i in seiner Bedeutung festgelegt.
  - **Beispiel:** Relation in Tabellendarstellung

Schema

ANGESTELLTER

Satztyp (Relation)

Ausprägungen

PNR	NAME	TAETIGKEIT	GEHALT	ALTER
496	PEINL	PFOERTNER	2100	63
497	KINZINGER	KOPIST	2800	25
498	MEYWEG	KALLIGRAPH	4500	56

➥ **DB-Schema**: vollständige Strukturbeschreibung (Metadaten) ist vor der Speicherung von Objekten zu spezifizieren und dem DBS bekannt zu machen

Datenbanksysteme (2)

- **Datenmodell / DBS-Schnittstelle**
  - Operationen zur Definition von Objekttypen (Beschreibung der Objekte)
    ➥ DB-Schema: Welche Objekte sollen in die DB gespeichert werden?
  - Operationen zum Aufsuchen und Verändern von Daten
    ➥ AW-Schnittstelle: Wie erzeugt, aktualisiert und findet man DB-Objekte?
  - Definition von Integritätsbedingungen (*Constraints*)
    ➥ Sicherung der Qualität: Was ist ein akzeptabler DB-Zustand?
  - Definition von Zugriffskontrollbedingungen
    ➥ Maßnahmen zum Datenschutz: Wer darf was?

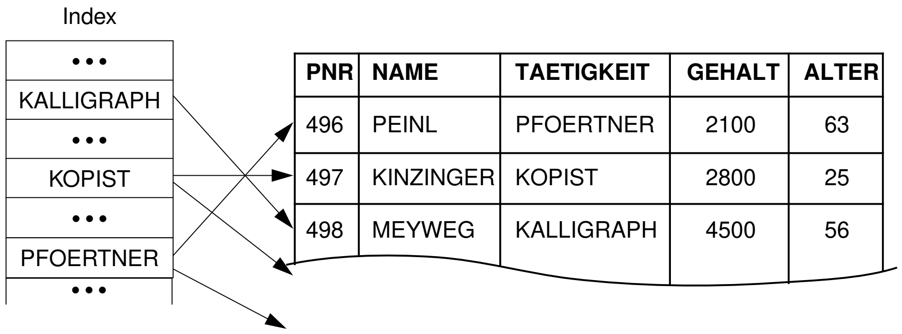
- **Art der DB-Sprache –** abhängig vom Datenmodell
  - formale Sprache
  - navigierend oder deskriptiv
  - satz- oder mengenorientiert
  - **Auswahlvermögen**: mindestens Prädikatenlogik erster Ordnung

- **Art der Suche**
  - Zeichen-/Wertvergleich:
    (TAETIGKEIT = ‘PFOERTNER’) AND (ALTER > 60)
  - <u>**exakte**</u> Fragebeantwortung:
    alle Sätze mit spezifizierter Eigenschaft werden gefunden
    (und nur solche)
  - Suche nach syntaktischer Ähnlichkeit: (TAETIGKEIT = ‘%PF%RTNER’)
    LIKE-Prädikat (in SQL) entspricht der Maskensuche

Datenbanksysteme (3)
- Suche bei strukturierten Daten

- **Sequentielle Suche?**
  - im Mehrbenutzerbetrieb: Jeder sucht nach anderen Informationen!
  - Antwortzeit??

- **Indexierung bei strukturierten Daten**
  - Invertierung von Attributen erlaubt „direkten Zugriff“ über die einzelnen Werte eine Attributs
  - Beispiel:
    ANGESTELLTER.TAETIGKEIT = {PFOERTNER, KOPIST, KALLIGRAPH, ...}



  - Invertierung eines Attributs wird bestimmt durch den erwarteten Leistungsgewinn bei der Anfrageauswertung
  - Mehrattribut-Indexierung (TAETIGKEIT | ALTER) erhöhen den Spezialisierungsgrad bei der Nutzung
  - Kombination von invertierten Attributen (TAETIGKEIT und ALTER)
    bei der Anfrageauswertung erlaubt verbesserte Nutzungsflexibilität
    und höheren Leistungsgewinn:
    (TAETIGKEIT = ‘PFOERTNER’) AND (ALTER = 50)

Beispiel: Relationenmodell und SQL

Schema

FB

FBNR	FBNAME	DEKAN

STUDENT

MATNR	SNAME	FBNR	STUDBEG

PRÜFUNG

PNR	MATNR	FACH	DATUM	NOTE

Ausprägungen

FB

FBNR	FBNAME	DEKAN
FB 9	WIRTSCHAFTSWISS	4711
FB 5	INFORMATIK	2223

STUDENT

MATNR	SNAME	FBNR	STUDBEG
123 766	COY	FB 9	1.10.95
225 332	MÜLLER	FB 5	15. 4.87
654 711	ABEL	FB 5	15.10.94
226 302	SCHULZE	FB 9	1.10.95
196 481	MAIER	FB 5	23.10.95
130 680	SCHMID	FB 9	1. 4.97

PRÜFUNG

PNR	MATNR	FACH	PDATUM	NOTE
5678	123 766	BWL	22.10.98	4
4711	123 766	OR	16. 1.98	3
1234	654 711	DV	17. 4.97	2
1234	123 766	DV	17. 4.97	4
6780	654 711	SP	19. 9.98	2
1234	196 481	DV	15.10.97	1
6780	196 481	BS	23.12.97	3

Q1: Finde alle Studenten aus FB 5, die ihr Studium vor 1995 begonnen haben.

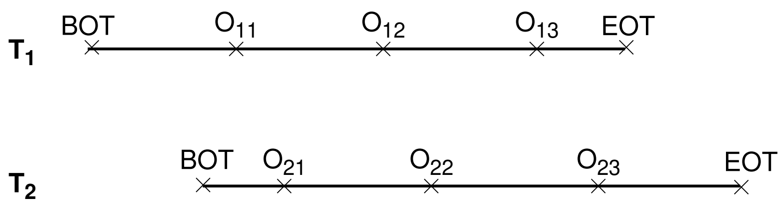
```
SELECT   *
FROM     STUDENT
WHERE    FBNR = 'FB5' AND STUDBEG < '1.1.95'
```

Q2: Finde alle Studenten des FB 5, die im Fach Datenverwaltung eine Note 2 oder besser erhalten haben.

```
SELECT   *
FROM     STUDENT
WHERE    FBNR = 'FB5' AND MATNR IN
         (SELECT   MATNR
          FROM     PRÜFUNG
          WHERE    FACH = 'DV' AND NOTE ≤ '2')
```


Datenbanksysteme (4)

- **Ablaufkontrollstruktur: Transaktion**



BOT: Begin of Transaction EOT (Commit): End of Transaction

O_{ij} : DB-Operation; Lese- und Schreiboperationen auf DB-Daten

mit den Eigenschaften

- **A**tomicity: Alles-oder-Nichts
- **C**onsistency: Gewährleistung der Integritätsbedingungen
- **I**solated Execution: „logischer Einbenutzerbetrieb“
- **D**urability: Persistenz aller Änderungen

- **Einsatzformen**
  - zentralisiertes DBS, Hauptspeicher-DBS
  - Mehrrechner-DBS (lokal oder ortsverteilt)
  - Hochleistungs-Transaktionssysteme
  - . . .

Information-Retrieval-Systeme (IRS)

- **Aufgaben/Eigenschaften von IRS**
  - Verwaltung von Dokumenten, Büchern, Abstracts usw.
  - effiziente Suche in großen Datenmengen
  - typischerweise nur Retrieval im Mehrbenutzerbetrieb
  - Anfragesprache für Retrieval
    ➥ Annäherung an natürliche Sprache erwünscht

- **Dokumente in IRS**
  - unstrukturierte Daten, keine dem IRS bekannte Dokumentstruktur
  - Beschreibung der Objekte durch Dokumenttyp und Wert (D_i/W_k)
  - Es gibt keine nähere Beschreibung oder Spezifikation von Struktur und Semantik, die W_k in seiner Bedeutung festlegt

    ➥ IRS verwaltet „lange“ Werte (z. B. Texte eines Abschnitts, Kapitels oder Buches) und stellt dafür Container (verschiedenen Typs) zur Verfügung

- **Beispiel**
  The two day workshop will be held in Edinburgh, the historic capital city of Scotland, between the Interact and VLDB conferences, thereby allowing attendees at either of these major international conferences to attend the workshop, and vice versa. UIDIS will be a limited numbers workshop, to encourage an informal atmosphere and to provide plenty scope for discussion and debate as well as presentations and demonstrations.

- **Indexierung bei unstrukturierten Daten**
  - Invertierung des gesamten Textes eines Dokumentes D_i
  - Wie wird inhaltsbasiert (nach W_k) gesucht?
  - Was sind geeignete Schlüsselwörter?

Information-Retrieval-Systeme (2)

- **Art der Suche**
  - Anfragen relativ **unscharf**
  - **Mehrdeutigkeit:**
    bei Wörtern in Dokumenten und Anfragen
    (Synonym-, Homonymproblem usw.)
  - Ähnlichkeitssuche (nearest neighbor, best match, pattern matching usw.)
  - Ergebnisbewertung: Relevanzproblem (Precision, Recall)

- **Verbesserung**
  - Synonymsuche, unscharfe Suche usw. sollen unterstützt werden
  - Einsatz eines Thesaurus (komplex organisiertes Wörterbuch):
    Festlegung von Beziehungen zwischen Begriffen

    ➥ „Verstehen natürlicher Sprache“, Erkennen von Mehrdeutigkeiten sind **„harte Probleme“**
    Bsp.: “Time flies like an arrow –
    fruit flies like a banana” (Groucho Marx)

- **Einsatzmöglichkeiten**
  - Bibliotheken
  - Literaturrecherche, z. B. im Internet (WWW)
  - Informationsdienste, Informations-Broker:
    Chemie, Recht, Patente, . . .

Semi-strukturierte Dokumente

- **HTML-Dokumente im WWW**
  - Aufbau des Dokumentes (syntaktische Struktur) ist festgelegt
  - Formatierungsanweisungen (Tags) lassen grobe Rückschlüsse auf den Inhalt des Dokumentes zu: TITLE, HREF, ADDRESS, ...
  - Inhalt des Dokumentes ist jedoch nicht weiter beschrieben; es gibt keine Metadaten, die die Bedeutung genauer festlegen

➥ WWW-Browser kann HTML-Dokumente aufbereiten und graphisch darstellen, ohne den Inhalt zu kennen

- **Beispiel: HTML-Dokument, semi-strukturiert**

```
<!DOCTYPE HTML PUBLIC „-//W3C//DTD HTML 3.2//EN“>
<HTML>
<HEAD>
  <TITLE>Publications 1998</TITLE>
  <META NAME=“GENERATOR“ CONTENT=“Mozilla/3.01Gold (X11; I; SunOS 4.1.3 sun4m) [Netscape]“>
</HEAD>
<BODY BACKGROUND=“http://www.uni-kl.de/AG-Haerder/pics/paper.jpg“>
<P><BASE HREF=“http://www.uni-kl.de/AG-Haerder/publications/“><A HREF=
“http://www.uni-kl.de/AG-Haerder/“ target=“_top“>[Top]</A>
<A HREF=“publications.html“>[Up]</A> <A HREF=“p1997.html“>[Next]</A> [Deutsche Version]<BR>
<H1 ALIGN=CENTER>Publications 1998</H1>
<HR NOSHADE></P>
<H3>Last update: 11/30/98 </H3>
<P>De&szlig;loch, S., H&auml;rder, T., Mattos, N., Mitschang, B., Thomas, J.:<BR>
<A HREF=“http://www-agdvs.informatik.uni-kl.de:18070/publications/DHMMT98.VLDB.html“>Advanced Data
Processing in KRISYS: Modeling Concepts, Implementation Techniques, and Client/Server Issues</A>,
in: VLDB Journal 7:2, 1998, pp. 79-95.</P>
<ADDRESS><A HREF=“mailto:wwwhaerder@informatik.uni-kl.de“>wwwhaerder@informatik.uni-kl.de</A>
</ADDRESS>
</BODY>
</HTML>
```

➥ **Indexierung und Suche** erfordern DBS- und IRS-Techniken

Semi-strukturierte Dokumente (2)

- **Formatierungssprachen dienen dem Austausch von Dokumenten**
  - Es gibt eine Vielzahl von Formatierungsprachen, die alle den internationalen Standard zur Textverarbeitung SGML als Meta-Sprache benutzen, um ihre Formate und Grammatik zu definieren

- **HTML**
  - ist eine Sprache zur Formatierung (Strukturierung) von Dokumenten (Texten) (HyperText Markup Language, Tag Language)
  - bietet eine vorgegebene Menge von Begrenzungs- und Formatierungs-anweisungen (>200) mit standardisierter Bedeutung
  - **Beispiel**

```
<H2>Second-Level heading </H2>
<P>This is a passage of text that probably belongs to the heading
immediately above </P>
```

  - vermischt Strukturierungs- und Darstellungsaufgaben
  - kann die Suche von Dokumenten kaum unterstützen

- **DocBook**
  - weitere Sprache zur Textformatierung (Software-Dokumentation)
  - **Beispiel**

```
<SECT2>
<TITLE>Second-level heading </TITLE>
<PARA> This is a passage of text that certainly belongs to heading above.
We know this because both are contained in the same SECT2 element.
</PARA>
</SECT2>
```

  ➥ Jede Sprache ist auf eine bestimmte Kategorie von Dokumenten zuge-schnitten

Semi-strukturierte Dokumente (3)

- **Neue Anforderungen**
  - **Erweiterbarkeit**, um nach Bedarf neue „Tags“ zu definieren
  - **Struktur**, um komplexe Daten zu modellieren und abzubilden
  - **Validierung**, um die strukturelle Korrektheit der Daten zu überprüfen
  - **Medienunabhängigkeit**, um Inhalte in verschiedenen Formaten zu publizieren
  - **Hersteller- und Plattformunabhängigkeit**, um entsprechende Dokumente mit standardisierter SW zu verarbeiten

- **Neuer Ansatz: XML (Extensible Markup Language)**
  - HTML4.0 $\in$ XML $\subset$ SGML
  - XML ist eine Metasprache, die Definition von Tags (in Anwendungskontexten) erlaubt
  - XML-Tags haben keine vordefinierte Semantik (wie bei HTML)

- **XML**
  - ist eine vereinfachte Form von SGML (und nicht eine erweiterte Form von HTML). Es wurden einige (für Web-Browser schwierige) Konstrukte von SGML weggelassen.
  - erlaubt die Definition einer beliebigen Anzahl von Formatierungssprachen für verschiedene Zwecke (Kategorien von Dokumenten) (z. B. Molekülstrukturen, japanische Texte, 3D-Objekte usw.)
  - dient zur Beschreibung von Struktur und Inhalt von Dokumenten (selbstbeschreibend)
  - kann als Sprache (Modell) zur Darstellung und zum Austausch von Dokumenten aufgefaßt werden

➥ Damit wird der **Austausch von Dokumenten** zwischen Programmen möglich.

➥ **Indexierung und Suche** werden drastisch verbessert, da XML Inhalte genauer spezifizieren kann.

Semi-strukturierte Dokumente (4)

- **Ziele**
  - XML kann für Daten erreichen, was Java für Programme bietet: Unabhängigkeit von Plattform (und Hersteller)

    ➥ Austausch von **Daten <u>und</u> Metadaten**
  - XML soll universelles, medienunabhängiges Publikationsformat bieten (für alle Klassen von Benutzern und alle Sprachen)

    ➥ „a single, completely internationalized format of almost unlimited power for both print and online publishing that is fully interoperable across all products and platforms“

- **Beispiel: wohl-geformtes XML-Dokument**

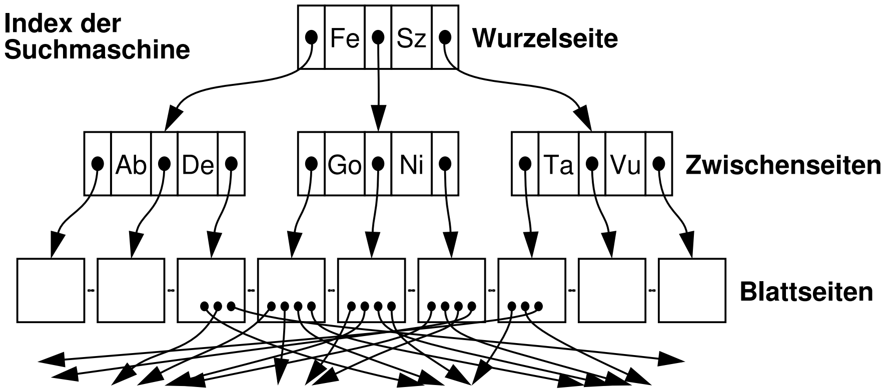
```
<book ISBN="1575213346">
    <title>Presenting XML</title>
    <author><firstname>Richard</firstname><lastname>Light</lastname></author>
    <author><firstname>Tim</firstname><lastname>Bray</lastname></author>
    <date>Sept.1997</date>
    <price currency="USD">19.99</price>
    <comment rating="4">
        <writtenby>
            <firstname>Harald</firstname><lastname>Schöning</lastname>
        </writtenby>
        <text>A quite useful book for <userclass>beginners</userclass>.</text>
    </comment>
    <comment>
        <writtenby>
            <firstname>A</firstname><lastname>Reader</lastname>
        </writtenby>
        <text> I did not like the cover</text>
    </comment>
</book>
```

➥ Die **Semantik** der Tags muß speziell festgelegt werden, z.B. durch Programme, Skripts oder deklarative Anweisungen für Formatvorlagen (style sheets)

Indexierung – Techniken

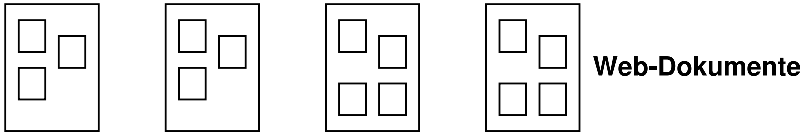
- **Index als B*-Baum**
  - organisiert alle Schlüssel (Schlagwörter) mit den zugehörigen URL-Listen in den Blättern
  - bietet direkten und sortiert-sequentiellen Zugriff auf alle Schlüssel über die Baumstruktur
  - erlaubt beliebiges Wachstum (dynamische Reorganisation) und garantiert stets eine ausgeglichene Baumstruktur (Höhenbalancierung)
  - besitzt einen sehr großen Verzweigungsgrad (fan-out: k+1 bis 2k+1 mit k>200) und eine geringe Höhe h, die logarithmisch mit der Anzahl N der Schlüssel wächst ($h = O(\log_{k+1} N)$)

- **Vereinfachtes Beispiel (k=1)**



Ergebnis:
Liste von URLs (http://...), ggf. aufbereitet durch Bewertungsfunktionen

Benutzerzugriff aufs WWW:



Indexierung – Probleme

- **Trennung, Indexierung, Klassifizierung erfordern Wortstamm-Zerlegung, Silbenbestimmung etc.**

- **Automatische Indexierung**
  - erfordert „**Verstehen natürlicher Sprache**“
  - erzeugt manchmal „köstlichste und kindlichste“ Fehler
  - bringt zuweilen „herrlichen Sprachquatsch mit Tiefsinn“ hervor

- **Beispiele:**

  Winterspor - torte
  Gra - binschrift
  Leich - tathleten
  Gebirg - sorte
  Bet - trost
  Schu - labschluß

  Beat - mung
  Fre - aks
  So - und
  Seinein - sel
  bein - halten
  Mattsch - eibe

- **Poetische Varianten:**

  Wel - traum
  Ka - minnische
  Auslöseele - ment
  Autoren - nen
  Seele - opard

- **Kuriose Trennungen:**

  Gassi - cherung
  Nachteil - zug
  Galauni - form
  Talent - wässerung
  Spargel - der
  Gehörner - ven
  Urin - stinkt

Ergebnis der Anfrageauswertung

- **Anfragen in relationalen DBS**

DATENSTRUKTUREN

Tabelle: PROFESSOREN

PNR	NAME	FB	FG
1234	HÄRDER	INFORMATIK	DBS
5678	STREICH	R&U	GIS
6780	MITSCHANG	INFORMATIK	DBS

Tabelle: MITARBEITER

MANR	NAME	FB	LEHRE
123 766	ZHANG	INFORMATIK	NEIN
130 680	RITTER	R&U	JA
196 481	ZIMMER	W-ING.	JA
225 332	STEIERT	R&U	NEIN
330 129	JAEDICKE	INFORMATIK	JA

SQL- ANFRAGE:

```
SELECT   NAME, FB
FROM     MITARBEITER
WHERE    LEHRE = 'JA'
```

ERGEBNIS:

NAME	FB
RITTER	R&U
ZIMMER	W-ING.
JAEDICKE	INFORMATIK

- genau festgelegter Bereich der Anfrage (Tabellen der FROM-Klausel)
- Verknüpfung verschiedener Tabellen ausschließlich über Werte
- Auswahl von Tupeln mit Hilfe von Prädikaten (WHERE-Klausel)

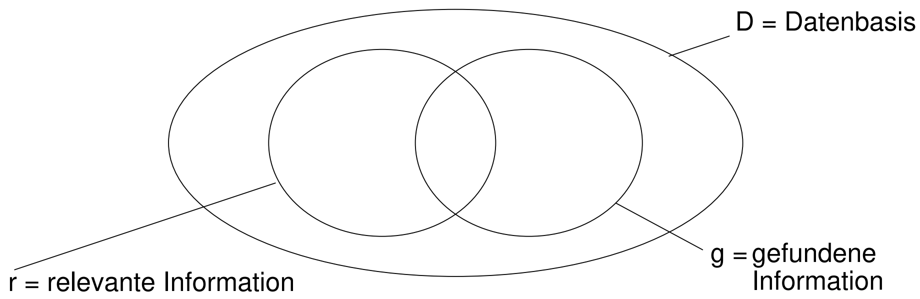
  ➥ Ergebnis liefert genau alle das Suchprädikat erfüllenden Tupel

- **Suche im WWW**
  - Einstieg über eine bekannte Adresse (URL) und Verfolgen von HyperLinks (Navigation)
  - Nutzung von Suchmaschinen
    - Zugriff über HTML-Formular
    - Angabe von Suchkriterien

      ➥ inhaltliche Suche von „außen“
    - Ergebnis: Liste von Adressen der WWW-Dokumenten (manchmal nach „Relevanz“ bewertet und geordnet)

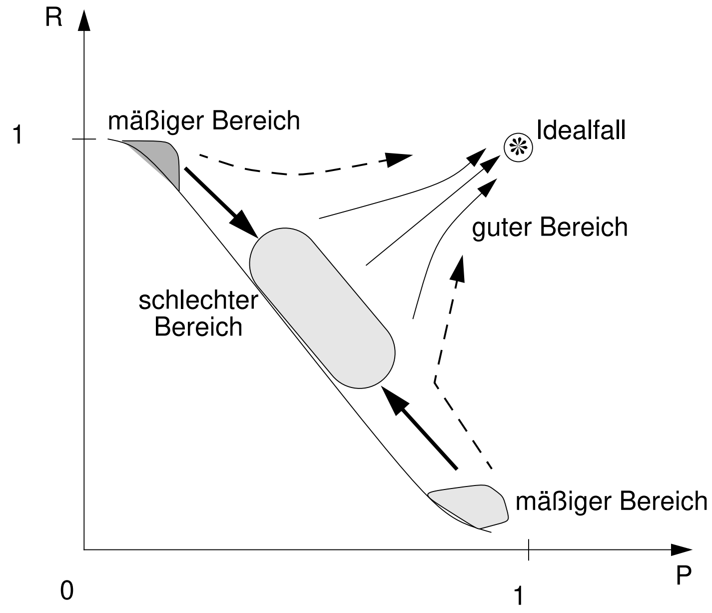
Ergebnis der Anfrageauswertung (2)

- **Auswertung von Anfragen in IRS (und analog im Web)**
  Recall R bezieht sich auf die relevanten Informationen
  Precision P bezieht sich auf die gefundenen Informationen



$$R = \frac{r \cap g}{r} \qquad P = \frac{r \cap g}{g}$$

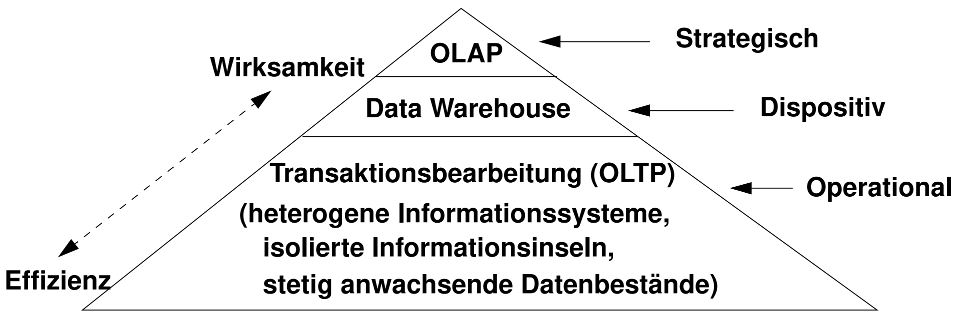
- **Messung der Systemgüte**
  empirische Ermittlung des R/P-Verhältnisses über eine Menge von Anfragen an das IRS



Klassen von DB-Anwendungen

- **Terminologie**
  - OLTP (*On-Line Transaction Processing*)
  - DW (*Data Warehouse*)
  - OLAP (*On-Line Analytical Processing*):
    Analyse betrieblicher Datenbestände
    - ROLAP (*Relational OLAP*), MOLAP (*Multi-dimensional OLAP*)
    - viele Anwendungsfelder von „Business Intelligence“
  - DSS (*Decision Support System*)
  - Data Mining: Aufspüren von inhärenten Daten-/Informationsmustern aus großen dynamischen Datenbeständen
    “In Data Mining applications, not only does the system define the semantics, it actually defines the queries. The user simply says ‘Go’, and the system produces what it believes to be useful answers.”
  - KDD (*Knowledge Discovery in Databases*), oft synonym zu Data Mining

- **Informationspyramide**



  - Data Warehouse als themenorientierte, integrierte, zeitlich veränderliche, nicht-flüchtige Datensammlung
  - Trennung der operationalen und informativen Daten

Klassen von DB-Anwendungen (2)

- **Beispiele für hohe Leistungsanforderungen**

1. Bankanwendungen/ Reservierungssysteme

Kontenbuchungen oder Platzreservierungen sollen mit einem Durchsatz von mehreren 1000 TPS und einer Antwortzeit < 1 sec bearbeitet werden.

2. Telefonvermittlung

Pro Telefongespräch ist ein Benutzerprofil aus der DB zu lesen sowie ein Abrechnungssatz zu schreiben. In Zeiten hohen Verkehrsaufkommens ist mit > 15.000 solcher Transaktionen pro Sekunde zu rechnen; die Antwortzeit sollte < 0.2 sec sein.

3. Management-Informationssysteme

Auf einer 500 GB großen DB sollen komplexe Ad-hoc-Anfragen ablaufen, die im ‘worst-case’ ein vollständiges Durchlesen der DB erfordern. Das DBVS soll einen Durchsatz von 5 TPS und eine Antwortzeit < 30 sec erreichen.

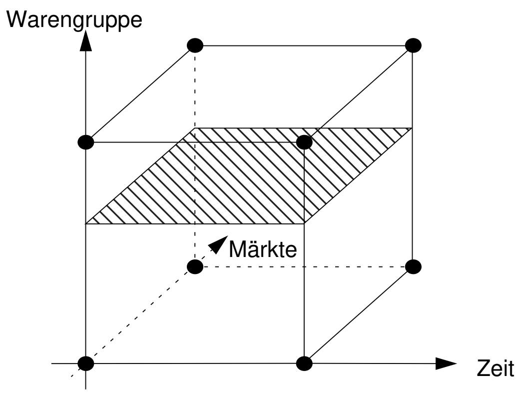
4. Aktienhandel (*online stock trading systems*)

- Broadcast von aktuellen Aktienkursen
  - gleichzeitige Benachrichtigung aller Makler
  - Makler können Selektionsprofil vorgeben
  - Makler können Angaben in PCs/Workstations puffern
- Bidding Service
  Entgegennahme von Angeboten zum Aktienankauf und -verkauf
- automatische Abwicklung von ‘Deals’
  - Durchsuchen der DB auf passendes Gegenangebot
  - Aktualisierung der Aktienstände,
  - Benachrichtigung der beteiligten Makler

Was sind Data-Warehouse-Systeme ?

Die Zielvorgabe für ein „Data Warehouse“ ist es, die im Unternehmen vorhandenen (und eventuell noch aufzubauende) Datenbestände dem Endbenutzer so bereitzustellen, daß dieser nicht nur *einen* vorgegebenen Blickwinkel (durch Programme realisiert) auf diese Daten einnehmen kann. Das bedeutet, daß sowohl der Datenbestand selbst als auch die benutzten Werkzeuge flexibel genug sein müssen, um *alle* anfallenden Fragestellungen zu beantworten.

Ein oft dargestelltes Beispiel solcher Blickwinkel ist der Absatz von verschiedenen Warengruppen, in verschiedenen Märkten unter Berücksichtigung der Zeit. Es ergibt sich damit folgender 3-dimensionaler Datenbereich.



Damit können nun unterschiedliche Fragen direkt beantwortet werden:

Für den Marktleiter:
Wie entwickelt sich Warengruppe X in meinem Markt im Zeitraum [Anfang, Ende]?

Für den Warengruppenmanager:
Welche Absatzverteilung auf Märkte bezogen gibt es für meine Warengruppe im Zeitraum [Anfang, Ende] (dargestellte Ebene in der Abbildung)?

Für den Finanzvorstand:
Wie entwickelt sich das Umsatzergebnis (als Summe über alle Märkte und alle Warengruppen) über die Zeit?

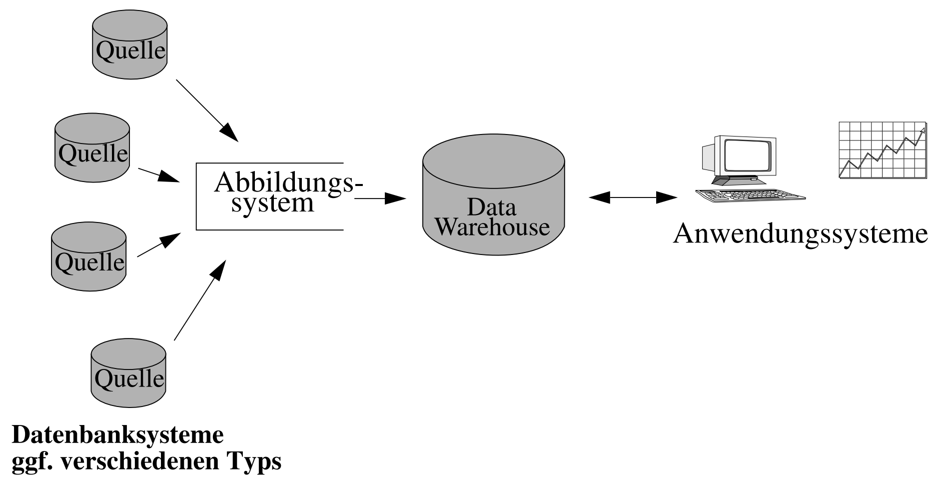
Die Analogie zum Warenhaus ist also dahingehend zu interpretieren, daß der Anwender durch die „Datenangebote“ geführt wird und die für ihn relevanten Informationen einfach „mitnehmen“ kann. Neben bereits dargestellten verschiedenen Blickwinkeln ergibt sich innerhalb der Dimensionen auch noch die Notwendigkeit einer Hierarchisierung: Beispielsweise kann die Warengruppendimension auf artikelgenaue Informationen verfeinert oder aber auf Sortimentsbereiche vergröbert werden.

Was sind Data-Warehouse-Systeme ? (2)

- **DW-Anwendungen**

Es sind betriebswirtschaftliche Kennzahlen, die sich als geeignete Managementunterstützung erwiesen haben, in sehr großen Datenbeständen (50 - 500 GB und mehr) abzuleiten und multidimensional aufzugliedern. Dazu ist eine Voraggregation der Daten, die in einem DW getrennt von den Daten der operativen DBS gehalten werden, aus Leistungsgründen unbedingt erforderlich. Weiterhin muß eine inkrementelle Aktualisierung dieser voraggregierten Daten aus den operationalen DBS (z. B. jede Nacht) erfolgen.

- **Grobaufbau eines DW-Systems (4 Systeme)**



- **Data Mining**

In DW oder operativen Datenbeständen sehr großer Volumina „schürfen intelligente Agenten“ selbständig nach impliziten Daten-/Informationsmustern, um bislang unbekannte Strukturen und Zusammenhänge aufzudecken. Solche für den Anwender interessanten Muster können

- Beziehungen zwischen Datensätzen oder zwischen Attributen eines Satzes
- gewisse Regelmäßigkeiten oder Regelabweichungen in Attributwerten

sein. Dazu ist das Erkennen von unscharfen oder probabilistischen Regeln nötig.

Was ist Transaktionsverarbeitung ?

- **Drei Aspekte:**
  - Mit einer Transaktion (TA) wird ein Vorgang einer Anwendung in einem Rechensystem abgewickelt.
    Ein solcher Vorgang bildet typischerweise einen **nicht-trivialen Arbeitsschritt** (*unit of work*) in betrieblichen Abläufen.
  - Eine (On-line) Transaktion ist die Ausführung eines Programmes, das mit Hilfe von Zugriffen auf eine **gemeinsam genutzte Datenbank (DB) eine Anwendungsfunktion** erfüllt.
  - Eine Transaktion ist eine **ununterbrechbare Folge von DB-Operationen**, welche die Datenbank von einem logisch konsistenten in einen logisch konsistenten Zustand überführt.

- **Beispiele:**
  - Überweisen eines Geldbetrages von Konto zu Konto
  - Platzreservierung für einen Flug
  - Bearbeiten einer Bestellung
  - Anmelden eines Autos
  - Abbuchen eines Tankbetrages
  - Abwickeln eines Telefonanrufes, . . .

- **Transaktionsprogramm „Kontenbuchung“:**

Read message(acctno, tellerno, branchno, delta) from Terminal;
BEGIN TRANSACTION
UPDATE ACCOUNT

SET balance = balance + *delta*
WHERE acct_no = *acctno* and balance >= *delta*

UPDATE TELLER

SET balance = balance + *delta*
WHERE teller_no = *tellerno*

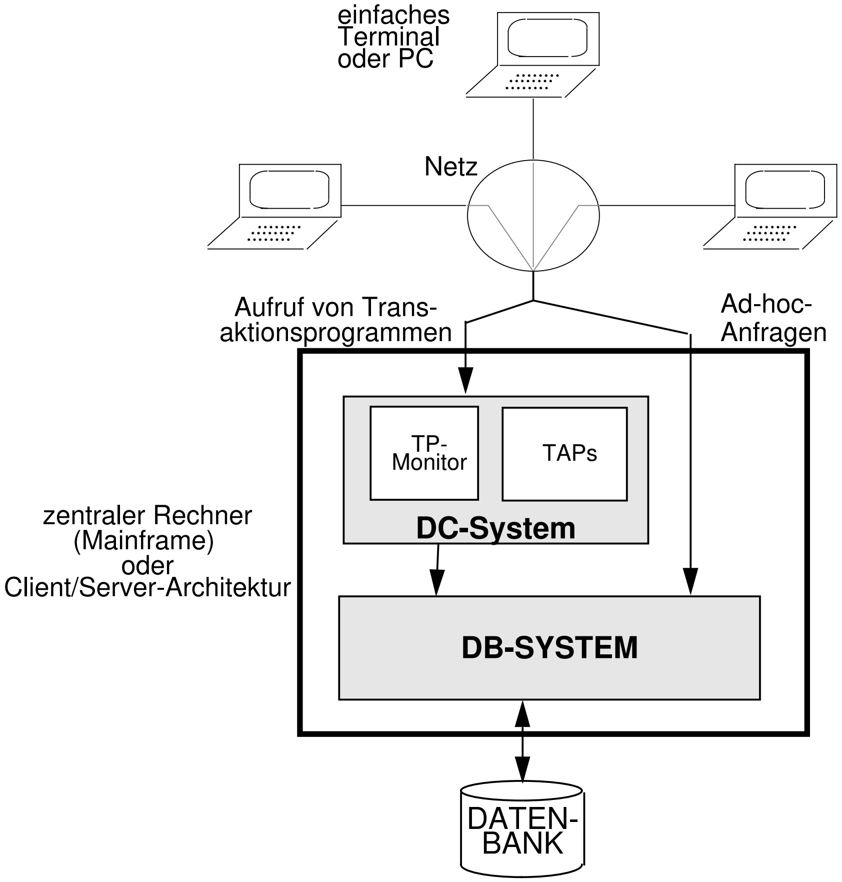
UPDATE BRANCH

SET balance = balance + *delta*
WHERE branch_no = *branchno*

INSERT INTO HISTORY (timestamp, values)
COMMIT TRANSACTION ;
Write message(acctno, balance, . . .) to Terminal

Einsatzszenarien von DBS

- **Wichtige Anwendungsklassen**
  - Transaktionssysteme
  - Ad-hoc-Anfragen
  - Data-Warehouse/Data-Mining (Schlagwort: Business Intelligence)

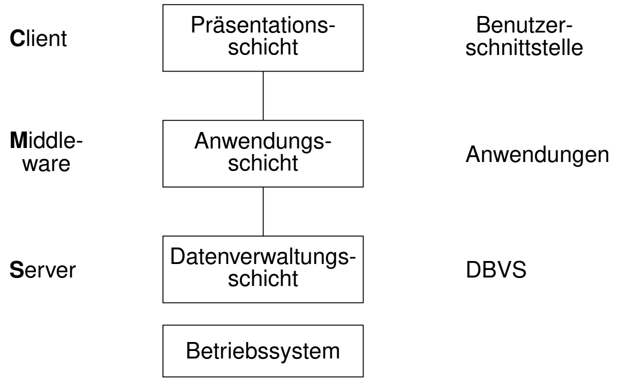


TAP	=	Transaktions-, Anwendungsprogramm
TP-Monitor	=	Transaction Processing Monitor
TPS	=	Transaction Processing System
DC-System	=	Datenkommunikationssystem

Einsatzszenarien von DBS (2)

- **Drei-Schichten-Anwendungsarchitektur**
  - 3-tier architecture
  - Client/Server-Prinzip

- **Allgemeines Modell**)



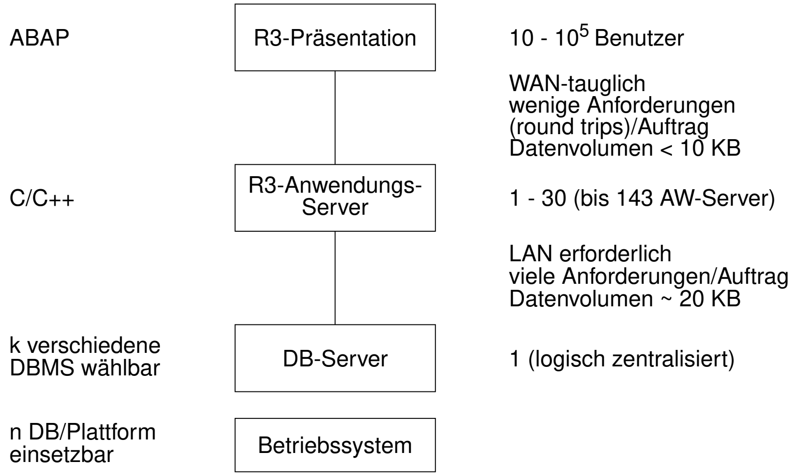
- **Client/Server-Architektur impliziert logische (funktionale) Aufteilung eines Systems**

- **Es sind verschiedene physische Aufteilungen (Abbildungen) möglich, z. B.**
  - **C + M + S** : 1 Rechner
  - **C** : n Rechner (PCs)
M + S : 1 Rechner
  - **C** : n
M : 1 .. k
S : 1 .. m

Einsatzszenarien von DBS (3)

- **Drei-Schichten-Anwendungsarchitektur am Beispiel von SAP/R3**

- **Realisierung in**



- **Funktionen in der mittleren Schicht (M)**
  - ABAP-Interpreter
  - TP-Monitor
  - AW-Systeme
  - Caching von DB-Daten
  - DBMS-Schnittstelle

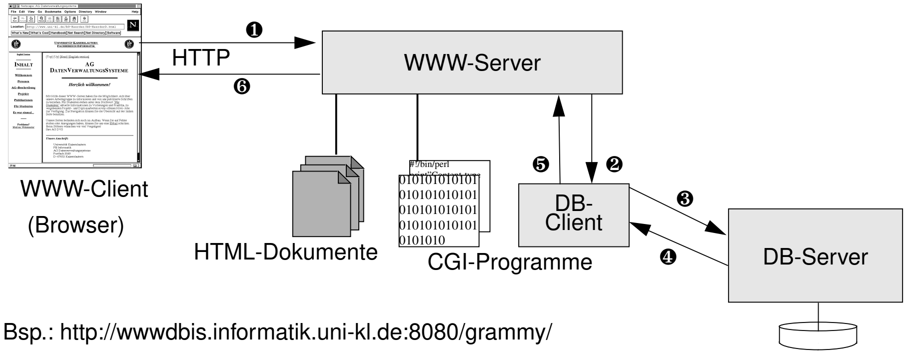
  → Unternehmensintegration über eine einzige DB, Skalierbarkeit sehr wichtig

- **Zahlen**
  - DB-Schema: > 15 000 Tabellen (Relationen)
200 000 Spalten (Attribute)
  - 35 M LOCs Anwendungsprogramme
  - Wachstumsrate pro größerer Version: + 30 %
  - Systemgröße (leer) auf Platte: 8 GB (foot print)
  - Unterstützung von mehr als 20 Sprachen

Einsatzszenarien von DBS (4)

- **DB-Zugriff übers Web**
  - Zustandslose Verbindung zwischen Browser und WWW-Server über HTTP (*HyperText Transfer Protocol*)
  - orts- und plattformunabhängiger Zugriff durch HTTP-Protokoll
  - DB-Zugriff durch externes CGI-Programm (*Common Gateway Interface*) bzw. Server-Erweiterungsmodul (*WWW Server API*) als DB-Client
  - HTML (*HyperText Markup Language*) dient zur Ergebnispräsentation

- **Anbindung über CGI (Beispiel)**

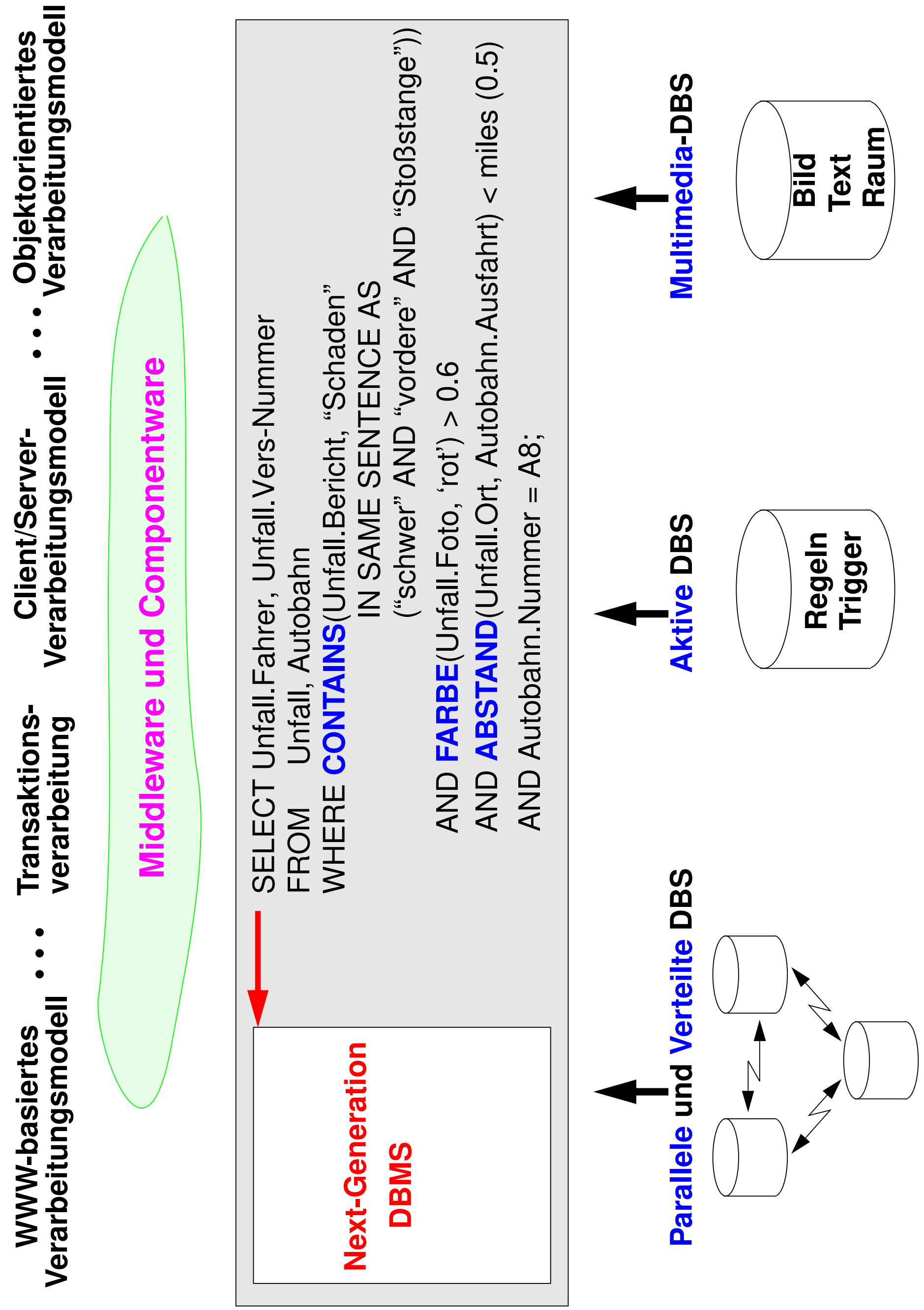


Bsp.: http://wwwdbis.informatik.uni-kl.de:8080/grammy/

- **Vorgehensweise:**
  ❶ Abschicken des Formulars mit aktuellen Parameterwerten
  ❷ Starten des entsprechenden CGI-Programms (DB-Client)
  ❸ Anfragen des CGI-Programms an den DB-Server
  ❹ Übertragung der Ergebnismenge zum DB-Client
  ❺ Rückgabe der erstellten HTML-Seite an den WWW-Server
  ❻ Übertragung der HTML-Seite zum Browser

- **Wie lassen sich DB-Server besser integrieren?**
  - einfachere Administration, neue Anfragemöglichkeiten
  - einfachere Anwendungsentwicklung
  - bessere Kontrolle von Integritätsbedingungen (Referentielle Integrität lokaler Hyperlinks), „intelligentes“ HTML

The Big Picture



Zusammenfassung

- **Transakionsparadigma**
  - macht weitreichende Zusicherungen für die Verarbeitung von DB-Daten
  - ACID-Eigenschaften müssen in einer Rechnerumgebung (aufwendig) nachgebildet werden
  - erlaubt die Implementierung von „Vertragsrecht“
- **Information und Informationssysteme**
  - Daten: objektive Welt der nicht-interpretierten Daten
  - Information: subjektive Welt der bewerteten Daten
  - Heterogenität, Wachstum, Anforderungsvielfalt u. a. führen oft auf unabhängige IS, die zusammen als kooperatives IS die angestrebte Leistung erbringen müssen

    ➥ „grob“: DBS + AWS = KIS
  - wichtige Anwendungsklassen für
    - operierende Ebene: OLTP
    - planende/kontrollierende Ebene: DW, OLAP
    - strategische Ebene: OLAP, DSS
- **Kernkomponenten rechnergestützter Informationssysteme**
  - Datenbanksystem
  - TP-Monitor (DC-System)
- **DBS-Technologie**
  - Verwaltung von persistenten Daten
  - effizienter Datenzugriff
  - flexibler Mehrbenutzerbetrieb
  - Konzepte:

    Datenmodell und DB-Sprache,

    Transaktion als Kontrollstruktur