DATENBANKANWENDUNG

Wintersemester 2013/2014

PD Dr. Holger Schwarz
Universität Stuttgart, IPVS
holger.schwarz@ipvs.uni-stuttgart.de

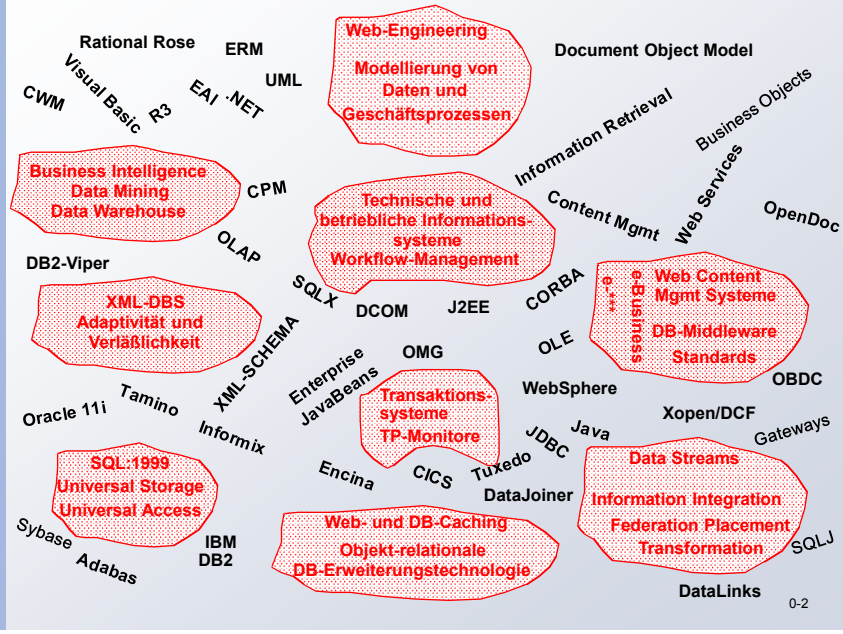
Beginn: **23.10.2013**
Mittwochs: **11.45 – 15.15 Uhr, Raum 46-268 (Pause 13.00 – 13.30)**
Donnerstags: **10.00 – 11.30 Uhr, Raum 46-268**
11.45 – 13.15 Uhr, Raum 46-260

http://wwwlgis.informatik.uni-kl.de/cms/courses/datenbankanwendung/

Die verwendeten Vorlesungsunterlagen basieren auf Vorlesungsunterlagen von Prof. Dr.-Ing. Dr. h. c. Theo Härder

TECHNISCHE UNIVERSITÄT KAISERSLAUTERN

Weltbild des Lehrgebietes Informationssysteme



Ziele

- **Vermittlung von Grundlagen- und Methodenwissen* zur Anwendung von Datenbanksystemen; Erwerb von Fähigkeiten und Fertigkeiten für DB-Administrator und DB-Anwendungs-entwickler**
  - Entwurf, Aufbau und Wartung von Datenbanken sowie Programmierung und Übersetzung von DB-Programmen, insbesondere auf der Basis von
    - Relationenmodell und SQL
    - objektorientierten und objekt-relationalen Datenmodellen mit Bezug auf die Standards ODMG und SQL:1999
  - Sicherung der DB-Daten und der Abläufe von DB-Programmen
    - Transaktionsverwaltung
    - Synchronisation
    - Fehlerbehandlung (Logging und Recovery)
    - Semantische Integrität, aktive DB-Mechanismen
    - Datenschutz und Zugriffskontrolle

* Grundlagenwissen ist hochgradig allgemeingültig und nicht von bestimmten Methoden abhängig. Die Halbwertzeit ist sehr hoch. Methodenwissen muss ständig an die aktuelle Entwicklung angepasst werden. In der Informatik haben sich die entscheidenden Methoden alle 8-10 Jahre erheblich geändert. Werkzeugwissen ist methodenabhängig. Werkzeuge haben in der Informatik oft nur eine Lebensdauer von 2-3 Jahren.

Ziele (2)

- **Voraussetzungen für Übernahme von Tätigkeiten ***
  - Entwicklung von datenbankgestützten Anwendungen
  - Nutzung von Datenbanken unter Verwendung von (interaktiven) Datenbanksprachen
  - Systemverantwortlicher für Datenbanksysteme, insbesondere Unternehmens-, Datenbank-, Anwendungs- und Datensicherungsadministrator

* "To bankrupt a fool, give him information." (Nassim Nicholas Taleb)
People make the mistake of stuffing their heads with words and numbers. They pile in more and more information hoping that sooner or later they'll have enough information to finally succeed at something.
But information only goes so far...and if you get too overloaded with it, you can go bankrupt learning to succeed instead of going out there and succeeding.

Übersicht

Übersicht (2)

Übersicht (3)

9. Logging und Recovery
- Fehlermodell und Recovery-Arten
- Logging-Strategien
- Recovery-Konzepte – Abhängigkeiten
- Sicherungspunkte
- Transaktions-, Crash- und Medien-Recovery

10. Integritätskontrolle und aktives Verhalten
- Semantische Integritätskontrolle
- Regelverarbeitung in DBS, Trigger-Konzept von SQL
- Definition und Ausführung von ECA-Regeln

11. Datenschutz und Zugriffskontrolle
- Technische Probleme des Datenschutzes
- Konzepte der Zugriffskontrolle, Zugriffskontrolle in SQL
- Sicherheitsprobleme in statistischen Datenbanken

12. Objektorientierung und Datenbanken (optional)
- Beschränkungen klassischer Datenmodelle
- Grundkonzepte der Objektorientierung
- SQL:1999 – Neue Funktionalität
  - ORDBS: Anforderungen, Architekturvorschläge
  - Erhöhung der Anfragemächtigkeit, Rekursion

13. Große Objekte (optional)
- Anforderungen und Verarbeitung mit SQL
- Lokator-Konzept, Speicherungsstrukturen, . . .

Literaturliste

Connolly, T., Begg, C.: Database Systems – A Practical Approach to Design, Implementation, and Management, 4th Edition, Addison Wesley, 2005

Elmasri, R., Navathe, S. B.: Grundlagen von Datenbanksystemen, Ausgabe Grundstudium, 3. Auflage, Pearson Studium, 2005

Härder, T., Rahm, E.: Datenbanksysteme – Konzepte und Techniken der Implementierung, Springer-Verlag, Berlin, 2001

Hoffer, J., Prescott, M., McFadden, F.: Modern Database Management (International Edition), 7. Auflage, Pearson Studium, 2005

Kemper, A., Eickler, A.: Datenbanksysteme – Eine Einführung, 6. Auflage, Oldenbourg-Verlag, 2006

Kifer, M., Bernstein, A., Lewis, P. M.: Database Systems – An Application-Oriented Approach, 2nd Edition, Pearson International Edition, 2006

Weikum, G., Vossen, G.: Transactional Information Systems, Morgan Kaufmann Publishers, San Francisco, CA, 2002

Literaturliste (2)

ZEITSCHRIFTEN

ACM TODS	Transactions on Database Systems, ACM Publikation (vierteljährlich)
Information Systems	Pergamon Press (6-mal jährlich)
The VLDB Journal	VLDB Foundation (vierteljährlich)
Informatik – Forschung und Entwicklung	Springer Verlag (vierteljährlich)
ACM Computing Surveys	ACM-Publikation (vierteljährlich)

TAGUNGSBÄNDE

SIGMOD	Tagungsbände, jährliche Konferenz der ACM Special Interest Group on Management of Data
VLDB	Tagungsbände, jährliche Konferenz „Very Large Data Bases"
ICDE	Tagungsbände, jährliche Konferenz „Int. Conf. on Data Engineering"
BTW	Tagungsbände der alle 2 Jahre stattfindenden Tagungen „Datenbanksysteme für Business, Technologie und Web" der GI, und weitere Tagungen innerhalb des GI-FB „DBIS"

und viele weitere Konferenzreihen

Notwendigkeit effizienter DBMS

- **Spektrum der Datenarten***
  - Nicht nur relationale Tabellen, sondern auch VITA-Daten (Video, Image, Text, Audio)**
  - Speicherung, Verwaltung, inhaltsorientierte Suche, Verknüpfung, …

* Bezeichnungen für Zehnerpotenzen: 3 kilo, 6 mega, 9 giga, 12 tera, 15 peta 18 exa, 21 zetta, 24 yotta und in der anderen Richtung:Bezeichnungen f 24 yecto, 21 zepto, 18 atto, 15 femto, 12 pico, 9 nano, 6 micro, 3 milli

** In 2010, we will cross the zettabyte barrier, weighing in at 1,200 exabytes http://www.catalystsecure.com/blog/2010/11/how-much-data-is-out-there-a-lot-more-than-you-might-think/

Notwendigkeit effizienter DBMS (2)

- „Datenbanktechnik ist eine nützliche Infrastruktur wie fließendes Wasser, das wir erst bemerken, wenn es fehlt".
- „Informationen sind in unserer vom Wettbewerb geprägten Welt ähnlich wie die Luft, die wir atmen – überall vorhanden und absolut lebenswichtig."
- **„We don't have better algorithms. We just have more data."** – Peter Norvig, Chief Scientist, Google

Entwicklung von Speichertechnologie und –bedarf?

- **Einige Trends**

Magnetic disks

Capacity	200+ GB	**x 10**	2+ TB
GB/$	0.05	**x 600**	30
IOPS	200	**x 1**	200

Solid state disks

Capacity	14 GB	**x 20**	256+ GB
GB/$	0.0003	**x 1,000**	0.5
IOPS (4KB read)	1,000 (SCSI)	**x 1,000**	1,000,000+ (PCIe)
			5,000+ (SATA)
IOPS (4KB write)	50 (SCSI)	**x 10,000**	500,000+ (PCIe+RAM)

Phase Change Memory

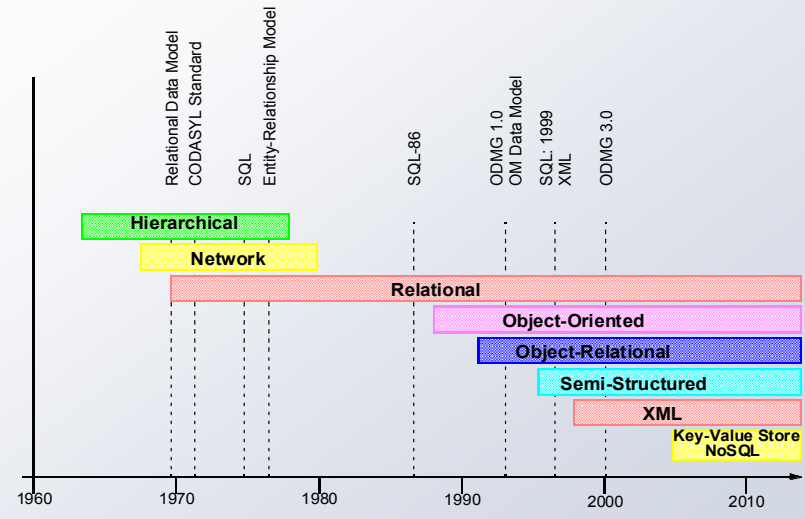
Capacity			1 GB chip (20-nm)
IOPS (64B read)			20,000,000+ (1 chip)
IOPS (64B write)			1,000,000+ (1 chip)

Entwicklung von Speichertechnologie und –bedarf? (2)

- **Racetrack technolgy (Storage Class Memory, SCM)***
  - SCM: Dramatische Reduktion von Platz und Energie bis 2020?
  - 1250 Racks (Gestellte) mit HDDs → 1 Rack mit SCM bei < 1/3 Energie
  - Längere Lebensdauer als Flash-SSDs
- **Petabyte Storage Device**
  - Entwurf für Archivspeicher: Lebensdauer > 50 Jahre
  - Keine Migration auf künftige Speichermedien erforderlich
  - „a petabyte in a rack unit"
- **Anwendungen mit „unersättlichem" Speicherbedarf**
  - Hollywood: Ein 3D-Film mit bis zu 1 PB
  - Gesundheitsfürsorge: heute 1TB pro Patient, künftig höhere Auflösung
  - Langzeit-Speicher wichtiger als SCM (Geschwindigkeit) für Industrie?

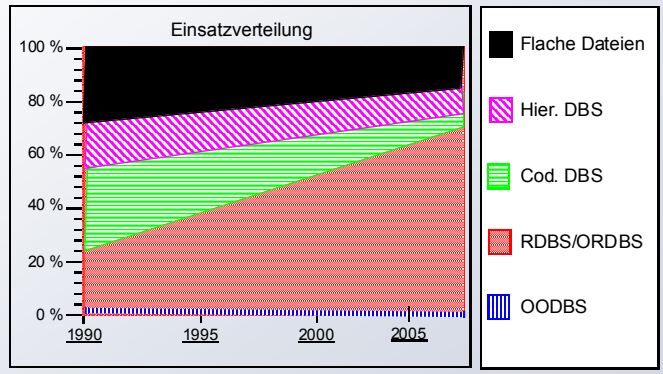
* http://en.wikipedia.org/wiki/Racetrack_memory: "This technology could enable a handheld device such as an MP3 player to store around 500,000 songs or around 3,500 movies --100 times more than is possible today -- with far lower cost and power consumption."

Evolution und Historie



Verteilung von DBS und Dateien

- **Es gibt verschiedenartige Datenmodelle und die sie realisierenden DBS**
  - relational und objekt-relational (RDBS/ORDBS)
  - hierarchisch (DBS nach dem Hierarchiemodell)
  - netzwerkartig (DBS nach dem Codasyl-Standard)
  - objektorientiert (OODBS

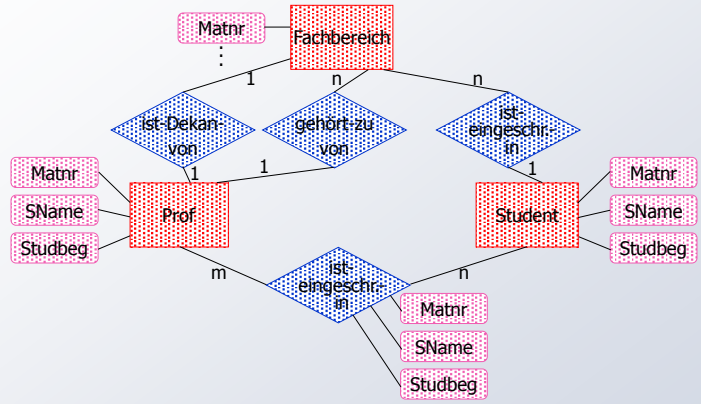


Verteilung von DBS und Dateien

- **Künftige DBS**
  - Aufstellung berücksichtigt nur strukturierte Daten. 85% der weltweit verfügbaren Daten aber sind semi- oder unstrukturiert (Internet, wiss. Aufzeichnungen und Experimente usw.)
  - SQL-XML-DBS, XML-SQL-DBS, native XML-DBS
- **Voraussetzung für die Vorlesung: Beherrschung von**
  - Informationsmodellen (erweitertes ER-Modell)
  - Relationenmodell und Relationenalgebra
  - SQL-92 als Standardsprache

Informationsmodellierung

- **Entity/Relationship-Diagramm einer Beispiel-Miniwelt**



- Spezifikation benutzerdefinierter Beziehungen (rein struktureller Natur)
- Klassifikation der Beziehungstypen (1:1, 1:n, n:m)

⇨ **relativ semantikarme Darstellung von Weltausschnitten**

- deshalb Verfeinerungen/Erweiterungen durch
  - **Kardinalitätsrestriktionen** ([1,1]:[1,10], [0:1]:[0:*])
  - **Abstraktionskonzepte** (Klassifikation/Instantiierung, Generalisierung/ Spezialisierung, Element-/Mengen-Assoziation, Element-/Komponenten-Aggregation)

Relationenmodell - Beispiel

- **DB-Schema**

FB

FBNR	FBNAME	DEKAN

PROF

PNR	PNAME	FBNR	FACHGEB

PROF

MATNR	SNAME	FBNR	STUDBEG

PRÜFUNG

PNR	MATNR	FACH	DATUM	NOTE

Relationenmdell – Beispiel (2)

- **Ausprägungen**

FB

FBNR	FBNAME	DEKAN
FB 9	WIRTSCHAFTSWISS	4711
FB 5	INFORMATIK	2223

PROF

PNR	PNAME	FBNR	FACHGEB
1234	HÄRDER	FB 5	DATENBANKSYSTEME
5678	WEDEKIND	FB 9	INFORMATIONSSYSTEME
4711	MÜLLER	FB 9	OPERATIONS RESEARCH
6780	NEHMER	FB 5	BETRIEBSSYSTEME

STUDENT

MATNR	SNAME	FBNR	STUDBEG
123 766	COY	FB 9	1.10.00
225 332	MÜLLER	FB 5	15.04.97
654 711	ABEL	FB 5	15.10.99
226 302	SCHULZE	FB 9	1.10.00
196 481	MAIER	FB 5	23.10.00
130 680	SCHMID	FB 9	1.04.02

PRÜFUNG

PNR	MATNR	FACH	PDATUM	NOTE
5678	123 766	BWL	22.10.03	4
4711	123 766	OR	16.01.02	3
1234	654 711	DV	17.04.03	2
1234	123 766	DV	17.04.03	4
6780	654 711	SP	19.09.03	2
1234	196 481	DV	15.10.03	1
6780	196 481	BS	23.10.03	3

Datenbank-anwendung

Ziele

Übersicht

Grundlagen

Ausblick

Relationenmodell – Beispiel (3)

- **Deskriptive DB-Sprachen**
  - hohes Auswahlvermögen und Mengenorientierung
  - leichte Erlernbarkeit auch für den DV-Laien
  - RM ist symmetrisches Datenmodell, d.h., es gibt keine bevorzugte Zugriffs- oder Auswertungsrichtung
- **Anfragebeispiele**

Q1: Finde alle Studenten aus Fachbereich 5, die ihr Studium vor 2000 begonnen haben

```
SELECT   *
FROM     STUDENT
WHERE    FBNR = 'FB5' AND STUDBEG < '1.1.00'
```

Q2: Finde alle Studenten des Fachbereichs 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')
```

Relationenmodell – Beispiel (4)

Q3: Finde die Durchschnittsnoten der DV-Prüfungen für alle Fachbereiche mit mehr als 1000 Studenten

```
SELECT    S.FBNR, AVG(P.NOTE)
FROM      PRÜFUNG P, STUDENT S
WHERE     P.FACH = 'DV' AND P.MATNR = S.MATNR
GROUP BY  S.NR
HAVING    (SELECT   COUNT(*)
           FROM     STUDENT T
           WHERE    T.FBNR = S.FBNR) > 1000
```


Bewertung – Relationenmodell*

- **Informationen des Benutzers**
  - ausschließlich durch den Inhalt der Daten
  - keine physischen Verbindungen, keine bedeutungsvolle Ordnung
- **Deskriptive Sprachen**
  - hohes Auswahlvermögen und Mengenorientierung
  - leichte Erlernbarkeit auch für den DV-Laien
- **Vorteile**
  - strenge theoretische Grundlage
  - einfache Informationsdarstellung durch Tabellen, keine Bindung an Zugriffspfade oder Speichertechnologie, keine Aussage über die Realisierung
  - hoher Grad an Datenunabhängigkeit
  - symmetrisches Datenmodell; d.h., es gibt keine bevorzugte Zugriffs- oder Auswertungsrichtung
  - Parallelisierung möglich, Verteilung der Daten über Prädikate
- **Nachteile**
  - zu starke Beschränkung der Modellierungsmöglichkeiten
  - schwerfällige und unnatürliche Modellierung bei komplexeren Objekten
  - Einsatz von nicht-prozeduralen Sprachen „soll" Ineffizienz implizieren!?
  - aber: Optimierung der Anforderungen liegt in der Systemverantwortung

* What's The Greatest Software Ever Written? By Charles Babcock, InformationWeek, Aug. 14, 2006
My No. 2 choice is IBM's System R, a research project at the company's Almaden Research Lab in San Jose, Calif., that gave rise to the relational database. In the 1970s, Edgar Codd looked at the math of set theory and conceived a way to apply it to data storage and retrieval. Sets are related elements that together make up an abstract whole. The set of colors blue, white, and red, for example, are related elements that together make up the colors of the French flag. A relational database, using set theory, can keep elements related without storing them in a separate and clearly labeled bin. It also can find all the elements of a set on an impromptu basis while knowing only one unique identifier about the set.
System R and all that flowed from it—DB2, Oracle, Microsoft SQL Server, Sybase, PostgreSQL, MySQL, and others— will have an impact that we're still just beginning to feel. Relational databases can both store data sets about customers and search other sets of data to find how particular customers shop. The data is entered into the database as it's acquired; the database finds relationships hidden in the data. The relational database and its SQL access language let us do something the human mind has found almost impossible: locate a broad set of related data without remembering much about its content, where it's stored, or how it's related. All that's needed is one piece of information, a primary key that allows access to the set. I like System R for its incredible smoothness of operation, its scalability, and its overwhelming usefulness to those who deal with masses of data. It's software with a rare air of mathematical truth about it.

The Big Picture

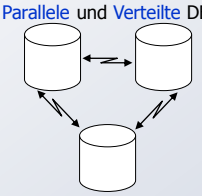
WWW-basiertes Verarbeitungs-modell . . . **Transaktions-verarbeitung** **Client/Server-Verarbeitungs-modell** . . . **Objektorientiertes Verarbeitungs-modell**

Middleware und Componentware

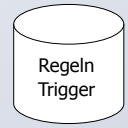
Next-Generation DBMS

```
SELECT  Unfall.Fahrer, Unfall.Vers-Nummer
FROM    Unfall, Autobahn
WHERE   CONTAINS(Unfall.Bericht, "Schaden"
                 IN SAME SENTENCE AS
                 ("schwer" AND "vordere" AND "Stoßstange"))
  AND   FARBE(Unfall.Foto, 'rot') > 0.6
  AND   ABSTAND(Unfall.Ort, Autobahn.Ausfahrt) < miles (0.5)
  AND Autobahn.Nummer = A8;
```

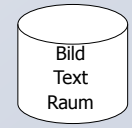
Parallele und Verteilte DBS



Aktive DBS



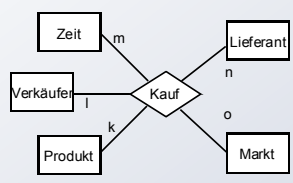
Multimedia-DBS



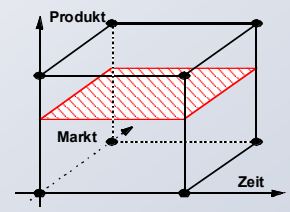
Was sind Data-Warehouse-Systeme?

Die Zielvorgabe für ein „Data Warehouse" ist es, die im Unternehmen vorhandenen (und eventuell noch aufzubauende) Datenbestände dem Endbenutzer so bereit-zustellen, 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 Produkten (Verkäufern und Lieferanten), in verschiedenen Märkten unter Berücksichtigung der Zeit.

ER-Schema mit 5 Dimensionen



3-dimensionaler Datenbereich



Was sind Data-Warehouse-Systeme?

Damit können nun unterschiedliche Fragen direkt beantwortet werden:

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

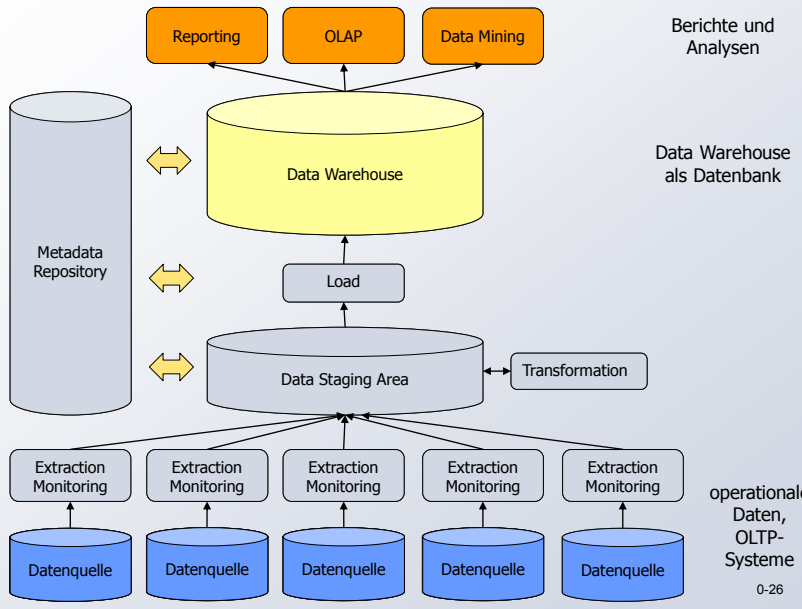
Für den Warengruppenmanager:
Welche Absatzverteilung auf Märkte bezogen gibt es für Produkt X im Zeitraum [Anfang, Ende] (dargestellte Ebene in der Abbildung)?

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 Produktdimension auf artikelgenaue Informationen verfeinert oder aber auf Warengruppen oder Sortimentsbereiche vergröbert werden.

Aktuelle Forschungs- und Entwicklungsthemen:
Real-time Data Analysis, Real-time Business Intelligence

Aufbau eines Data Warehouse

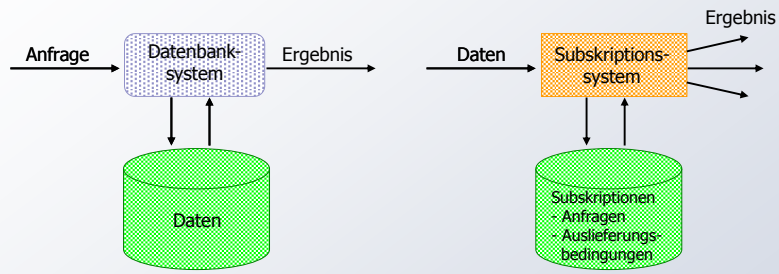


Begriffe

- Viele Namen, die nicht alle gleiche Bedeutung besitzen:
  - Data Mining, Knowledge Discovery, Business Intelligence, Data Exploration, Pattern Recognition, Information Retrieval, Knowledge Management, ...
- Dies ist unsere Sicht:
  - **Knowledge Discovery** ist ein Prozess zur Suche oder Erzeugung von Wissen aus großen Datenmengen. Seltener wird dazu auch der Begriff Data Exploration benutzt.
  - Eine Phase dieses Prozesses, Pattern Generation genannt, generiert relevante Informationen. In unserem Falle ist diese Phase gleichbedeutend mit **Data Mining**, hier kann aber auch z.B. Online Analytical Processing (OLAP) angesiedelt sein.
  - **Business Intelligence** bezeichnet den Einsatz von Knowledge Discovery im Unternehmen, um ökonomischen Nutzen zu erzielen.
  - Der Begriff **Pattern Recognition** wird eher in anderen Disziplinen wie KI (Bildverstehen) und Naturwissenschaften (z.B. bei Informationssystemen für Biochemie und Geographie (GIS)) verwendet, obwohl es sich hierbei im Prinzip auch um Knowledge Discovery handelt.
  - **Information Retrieval** ist gleichbedeutend mit Data Exploration, wird aber eher für textorientierte Informationen benutzt (Such-Maschinen, „Text-Mining").
  - **Knowledge Management** umfasst u. a. auch Knowledge Discovery. Im Vordergrund steht aber die Verwaltung von Wissen, damit es auch in Zukunft den Mitgliedern einer Organisation in geeigneter Weise (Geschwindigkeit, Qualität, Kosten, etc.) zur Verfügung steht.

Ein weiteres Paradigma – „Alles fließt" (panta rhei)*"

- **Vertauschte Rollen**



  - Statt Auswertung von gespeicherten Daten Filterung, Verknüpfung und Transformation von Datenströmen
  - Zentrale Bedeutung für die individuelle Informationsversorgung, insbesondere bei einer immer weiter fortschreitenden Verwendung vieler kleiner und damit mobiler Endgeräte

* Fälschlicherweise Heraklit zugeschriebene Formel für sein Weltbild

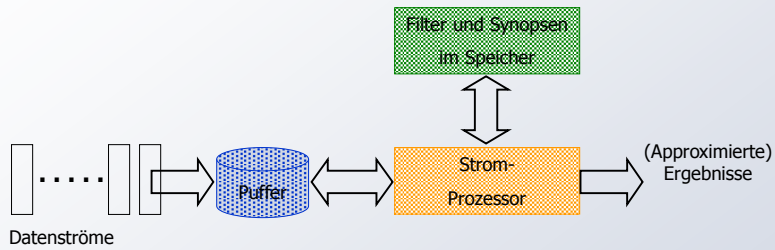
Ein weiteres Paradigma (2)

- **Wichtige Unterschiede**

Eigenschaft	Datenbankbasierte Informationssysteme	Subskriptionssysteme
Verarbeitungscharakteristik	Zustandsorientiert (globaler Zustand)	konsumorientiert (evtl. lokaler temporärer Zustand)
Anfragesemantik	isolierte Anfrage	Stehende Anfragen (‚standing query')
Zugriffscharakteristik	systemzentriert (‚row-set-model')	dokumentenzentriert (‚document-model')
Auswertungssemantik	komplexe Analyse	informativ, Auslöser detaillierter Analyse
Schemaaspekt	Existenz eines globalen Schemas	Zugriff auf lokale Schemata der partizipierenden Datenquellen

Datenströme - Auswertungsmodell

- **Daten strömen und sind nicht in einer DB gespeichert***

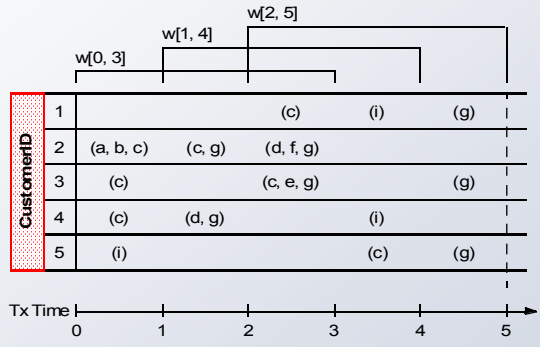


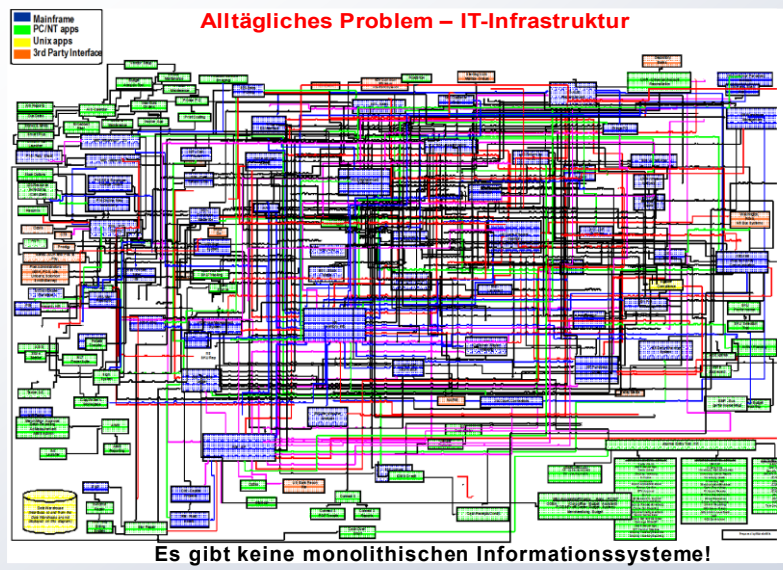
- Analyse von flachen Datenströmen (z.B. sensorgeneriert)
- Strombasierte Analyse von XML-Dokumenten
- Wie funktionieren Selektionen und Joins?

* Aktuelles Forschungsthema im Rahmen der dynamischen Informationsfusion

Datenströme – Auswertungsmodell (2)

- **Beispiel für einen Online-Transaktionsfluss**
  - Warenkorb-AW: Kunden kaufen Waren, die als Ströme zu analysieren sind
  - Bestimmung der Häufigkeit von Mustern (frequent pattern identification) ist offensichlich sehr schwierig!





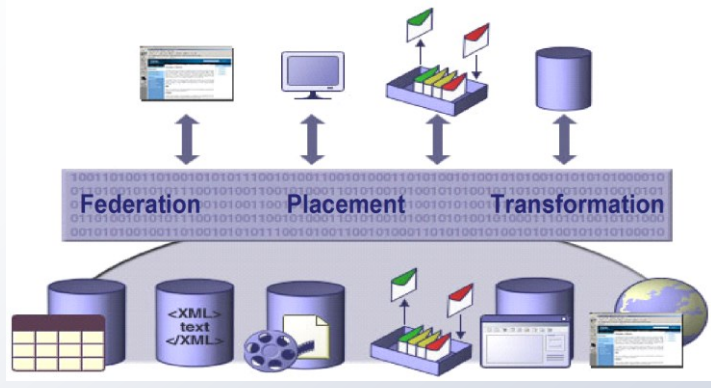
Verlässliche adaptive DBMS

- **DBMS ist nur eine (wichtige) Komponente in einem Informationssystem!**
- **Höherer Grad an „Selbst-Bewusstsein" gefordert**
  - Adaptivität hinsichtlich
    - Benutzer und Arbeitslasten
    - Betriebsmittel, Plattformen und Umgebungen
    - Information (Repräsentation, Inhalt, …)
- **Verhaltensmodelle im DBMS erforderlich**
  - in den Schichten
  - schichtenübergreifend – zusätzliche Kanäle für nicht-lokale Information
- **Adaptivität zwischen den Komponenten des Informationssystems**
  - heterogen und organisationsübergreifend
  - Agreement-Protokolle
- **Entwicklungsziele (J. Gray: 1998 Turing Lecture)***

1. Trouble-free systems: Build a system used by millions of people each day and yet administered and managed by a single part-time person.
2. Secure system: Assure that the system of problem 1 only services authorized users, service cannot be denied by unauthorized users, and information cannot be stolen (and prove it).
3. Assure that the system is unavailable for less than one second per hundred years – 8 9's of availability (and prove it).

* J. Gray is the recipient of the 1998 A. M. Turing Award. These problems, strongly related to database systems, are extracted from the text of the talk J. Gray gave in receipt of that award. http://www.research.microsoft.com/~gray/

Dynamische Informationsfusion



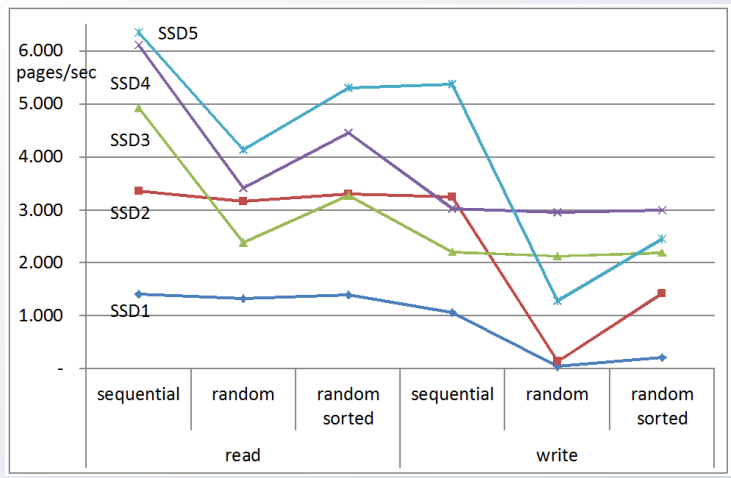
- **Paradigmenwechsel: bedarfsgetriebene, skalierbare Kopplung und Integration von Datenquellen, -strömen und –analyse-modellen (DSMs)**

Dynamische Informationsfusion (2)

- **DSMs**
  - Persistent gespeicherte Daten
  - Kontinuierliche Datenströme, in Zeitfenstern zu verarbeiten (Senesordaten, Web-Nutzerverhalten)
  - Abgeleitete DA-Modelle (data mining)
- **Skalierbarkeit**
  - Weltweiter Zugriff auf DSMs: Internet, Grid Computing, P2P, Web Services
  - Effiziente Auswahl der geeignetsten Komponenten aus einer großen und schnell wachsenden Menge
- **Dynamik**
  - Persisent gespeicherte Daten
  - Kontinuierliche Datenströme, in Zeitfenstern zu verarbeiten (Sensordaten, Web-Nutzerverhalten)
  - Abgeleitete DA-Modelle (data mining)
- **Datenqualität**
  - Daten variieren in Genauigkeit, Vollständigkeit, Aktualität und Dimensionalität
  - Zusätzliche Methoden und Metadaten

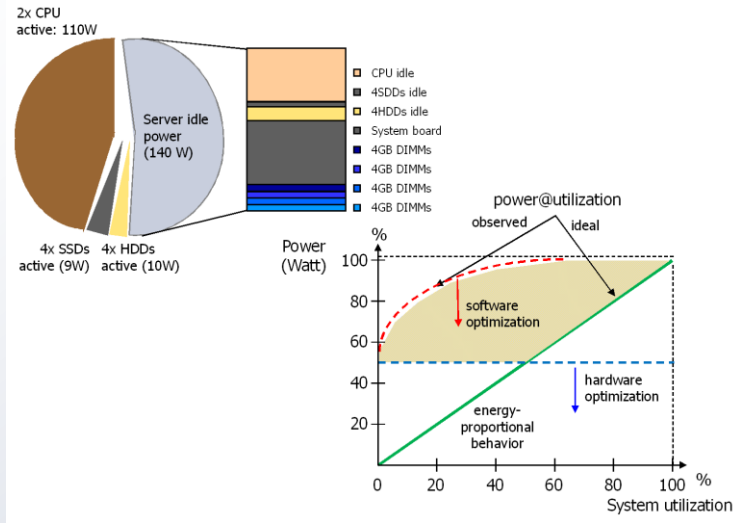
Energieeffizienz – Es müssen gigantische Datenmengen gespeichert und verarbeitet werden

- **IOPS von Flash-Speicher**



Energieeffizienz – Es müssen gigantische Datenmengen gespeichert und verarbeitet werden (2)

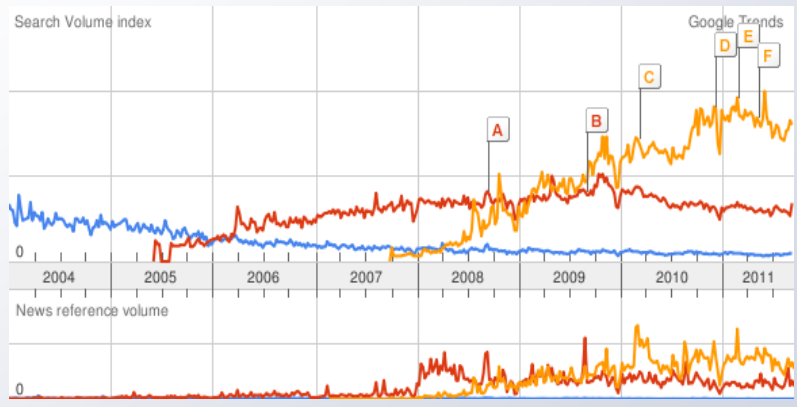
- **Neue Herausforderung: Energieproportionale Verarbeitung!**



Hype-Begriffe wechseln schnell!

- **Gemessen mit Google Trends**

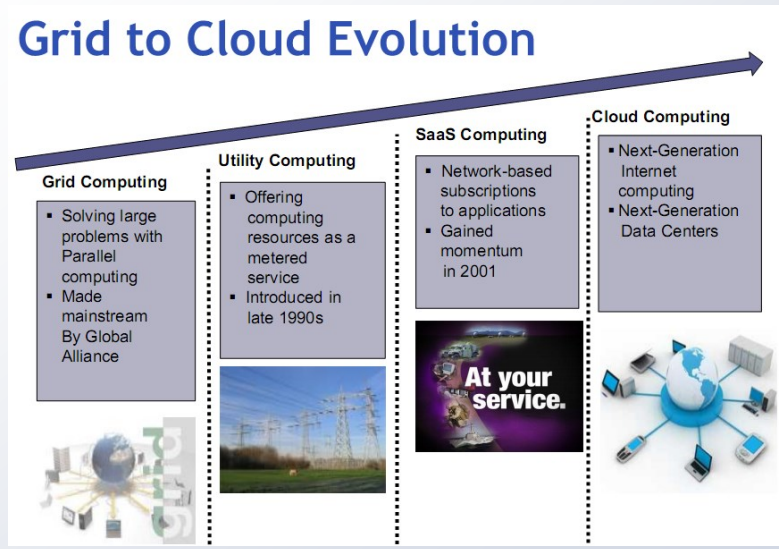
grid computing virtualization cloud computing



Hype-Begriffe wechseln schnell!

- **Was ist Grid Computing und was Cloud Computing?**
  - Flexible Zuordnung von Rechnern, Daten und Funktionen
  - Das macht auch „Seti@Home" (> 3 Mio Benutzer)! Was fehlt also noch?

Hype-Begriffe wechseln schnell! (2)



Pros and Cons of Cloud Computing (www.dsp-ip.com)

- **Pros**
  - Scale
  - Cost (CAPEX, OPEX)*
  - Advance architecture
  - Agility
  - Cost – Clouds are renowned for being cheap for storage and processing**
  - Elasticity – Growth and shrinkage
- **Cons**
  - Security & privacy (Is it safe? For whom and at what level?)
  - Regulatory compliance***: HIPPA, SOX, etc.
  - Interoperability and vendor lock-in
  - Lack of control
  - Standardization
- **Challenges**
  - Organizational barriers
  - Reliability (service outage)
  - Definition of SLAs (service-level agreements)
  - Service management (LCM), monitoring
  - Customization
  - Integration with other applications
  - Technology (limited languages & APIs)
- **Cloud Concerns**
  - ➥ Security is No. 1!

* Capex heißt capital expenditures und bedeutet Kapitaleinsatz. Opex heißt operating expenditure und bedeutet Instandhaltungsaufwand.
** Amazon main services: Elastic Cloud Service (EC2), Simple Storage Service (S3), Simple Database Service (SimpleDB), Simple Queue Service (SQS)
*** Handeln in Übereinstimmung mit geltenden Vorschriften

MyLifeBits

- **Selbst wenn der Mensch sein Leben lang kein Wort vergessen würde, ist die Speicherung der Informationsmenge heute schon leicht machbar***

Bilanz (durchschnittliche Zahlen für US-Bürger/Jahr)

	Stunden	Wort/Minute	Wörter/Jahr	MBytes
TV	1578	120	11 Mio.	50
Film	12	120	-	-
Lesen	354	300	6,4 Mio.	32

➥ ~6 GBytes von ASCII-Daten (Text) in 75 Jahren
Die automatische Erfassung dieser Informationen (Spracherkennung, und OCR oder elektronische Bücher/Zeitungen und ASCII-Skripte von TV-Sendungen) ist mit tragbaren Geräten möglich.

➥ Das gilt nicht für Ton, Bild und Bewegtbild!

* http://research.microsoft.com/barc/MediaPresence/MyLifeBits.aspx; http://en.wikipedia.org/wiki/MyLifeBits

MyLifeBits (2)

- **Aber in 20 Jahren?**
  Speicherbedarf für eine Person (nach Jim Gray)

Datentyp	Datenrate (Bytes/Sek.)	benötigter Speicher pro Stunde und Tag	benötigter Speicher für eine Lebenszeit
gelesener Text	50	200 KB; 2-10 MB	60-300 GBytes
gesproch. Text @ 120wpm	12	43 K; 0,5 MB	15 GBytes
Sprache (kompromiert)	1.000	3,6 MB; 40 MB	1,2 TBytes
Bewegtbild (komprimiert)	500.000	2 GB; 20 GB	1 PByte

  ➥ Aufzeichnung der gesamten Lebensgeschichte wird möglich

- **Weltweites Wachstum**
  - Professionelle Content Provider (Journalisten usw.): 2 GB Text pro Tag
  - Benutzergenerierte Web-Inhalte: dramatisch (siehe social networks)

Schlussfolgerungen

- Es wird genug Platten- und Bandspeicher geben, um alles zu speichern, was alle Menschen schreiben, sagen, tun oder photographieren.
  - Für das Schreiben gilt dies bereits heute
  - In einigen Jahren trifft das auch für die restlichen Informationen zu
  - Wie lange wird es noch dauern, bis alle VITA-Dokumente gespeichert werden?
- MyLifeBits: Aufzeichnung der gesamten Lebensgeschichte wird möglich*
- Rechner speichern und verwalten Informationen besser und effektiver als Menschen
  - Viele Platten und Kommunikationsverbindungen speichern direkt Informationen aus Rechner-zu-Rechner- und nicht mehr (nur) aus Mensch-zu-Mensch-Kommunikation
  - Wie lange wird es noch dauern, bis der Mensch die meiste gespeicherte Information gar nicht mehr zu sehen bekommt?
  - Wir müssen lernen, wie alles automatisch ausgewertet werden kann und was bei unserer knappen Zeit unserer besonderen Aufmerksamkeit bedarf.

* http://research.microsoft.com/barc/MediaPresence/MyLifeBits.aspx; http://en.wikipedia.org/wiki/MyLifeBits

Schlussfolgerungen (2)

- Künftige Entwicklung
  - Heute konzentriert man sich bei den „Digitalen Bibliotheken" auf die Eingabe: auf das Scanning, Komprimieren und OCT von Informationen.
  - Morgen wird anstelle der Eingabe die „relevante Auswahl" die wesentliche Rolle spielen: Selektion, Suche und Qualitätsbewertung von Informationen

  ➥ Wir können eine reale „World Encyclopedia" mit einem echten „planetary memory for all mankind" aufbauen, wie H.G. Wells bereits 1938 in seinem Buch „World Brain" geschrieben hat!