

*Data Everywhere – der lange Weg von
Datenansammlungen zu
Datenbanksystemen*

dbis

Datenbanken und Informationssysteme

Prof. Johann-Christoph Freytag, Ph.D.

Humboldt-Universität zu Berlin



Ringvorlesung SoSe 2005 – 19. Mai 2005

dbis

Agenda



Überblick

Plattentechnologie

Datenmodelle & Datenbanken

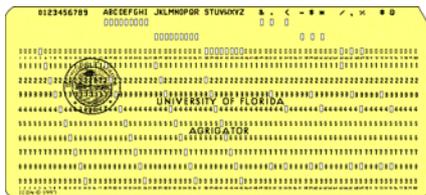
Relationales Technologie

Weitere Entwicklung

Prof. Johann-Christoph Freytag, Ph.D. – Data anywhere

2

- **Datensammlungen... und Speichertechnologie**
 - Herman Hollerith: „punch card tabulating machine“
 - Firmen – Vorläufer der IBM
 - Tabulating Machine Corp - 1896
 - Computing-Tabulating-Recording Company (C-T-R) - 1911
 - International Business Machines Corporation (IBM) - 1924



- **Erste Platte: RAMAC 350 der IBM – 1955/56**
 - „Random Access Method of Accounting and Control“
 - Entwickelt in San Jose, CA als sog. „bootleg“Projekt
 - Reynold B. Johnson (1906-1998): Technischer Leiter

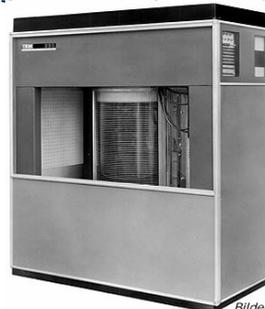


Bilder des Magnetic Disk Heritage Centers, Santa Clara, CA

RAMAC 350

- Technische Daten

- Gewicht: 1 Tonne
- Speicherkapazität: 5 MB (Wörter mit 7bits)
- 50 Scheiben – 24 inch (61cm) Durchmesser
- 1200 rpm – Plattenarm 200 µm über der Scheibe
- Speicherdichte: 100 bits per inch – 0,7 sec Zugriffszeit
- Kosten (1956): 50k\$ Miete (Vergl. Rolls Royce 10k\$ Kauf)

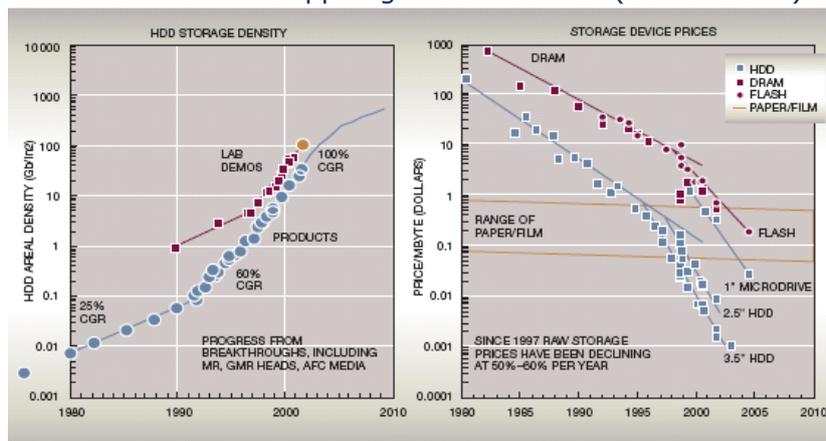


Bilder des IBM Archivs

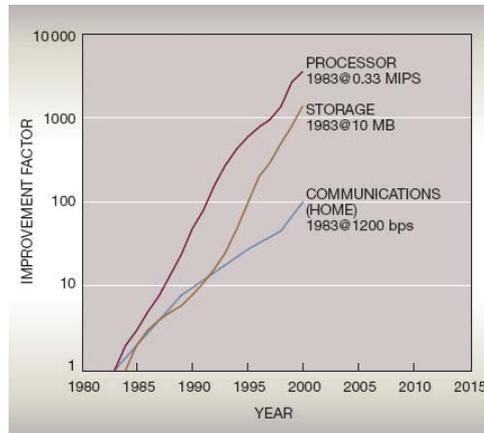
Plattenentwicklung

- Speicherdichte und Speicherpreis

- HDD Dichte: Verdopplung alle 18 Monate (Moore's Law)



Verbesserung in den letzten 25-30 Jahren



Quelle: IBM SYSTEMS JOURNAL, VOL 42, 206 NO 2, 2003

	1956	2005
Cost per megabyte	\$10,000	\$0.001
Areal density	1000 bits/in ²	80 GB/in ²
Access time	1 sec.	4 msec.

Quelle: Don Chamberlin, 2005

Ringvorlesung SoSe 2005 – 19. Mai 2005

dbis 

Agenda

Überblick

Plattentechnologie

Datenmodelle & Datenbanken

Relationales Technologie

Weitere Entwicklung

9

Prof. Johann-Christoph Freytag, Ph.D. – Data anywhere

Ringvorlesung SoSe 2005 – 19. Mai 2005

dbis 

Datenbanken – Historisches I

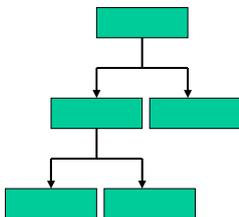
- Begriff Datenbanken (engl. databases)
 - Geprägt am Anfang der 60er Jahre
 - Erkenntnis: Information soll unabhängig von spezieller HW bzw. Maschine konzeptuell bearbeitet, strukturiert und manipuliert werden können
 - Ging einher mit Standardisierung von COBOL (Common Business Object Language) - 1960
 - Datendefinitions(sub)sprache
- Erstes Datenbankmanagementsystem 1961
 - Charles Bachman (General Electric Company)
 - Integrated Data Store (IDS)
 - Plattenbasiert mit Schemadefinition & Logging
 - Standardisierung durch Database Task Group (DBTG) - 1971:
 - CODASYL-Datenbanken (**C**onference on **D**ata **S**ystems **L**anguages)
Konferenz zwischen Militär, Wirtschaft & Computerherstellern (1959)
 - Möglich: Datenbanksysteme durch andere Firmen auf anderer HW

10

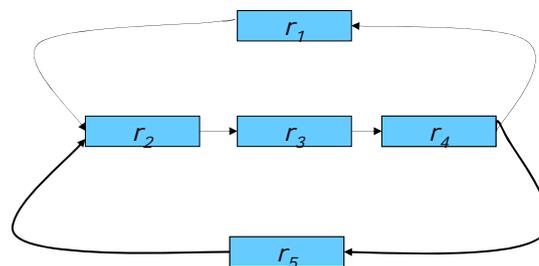
Prof. Johann-Christoph Freytag, Ph.D. – Data anywhere

- Parallel: Entwicklung der IBM: IMS
 - Information Management System (IMS) – seit 1968
 - Ursprünglich entwickelt mit der NASA für das Apollo Raumfahrtprogramm
 - Information Control System (zusammen mit Rockwell)
 - Nur auf IBM Rechnern verfügbar
 - Noch heute im Einsatz
 - Weiterentwickelt seit ca. 40 Jahren
 - Mehr als 1 Milliarde \$ Umsatz pro Jahr
 - Mehr Daten gespeichert als jedes andere DBMS

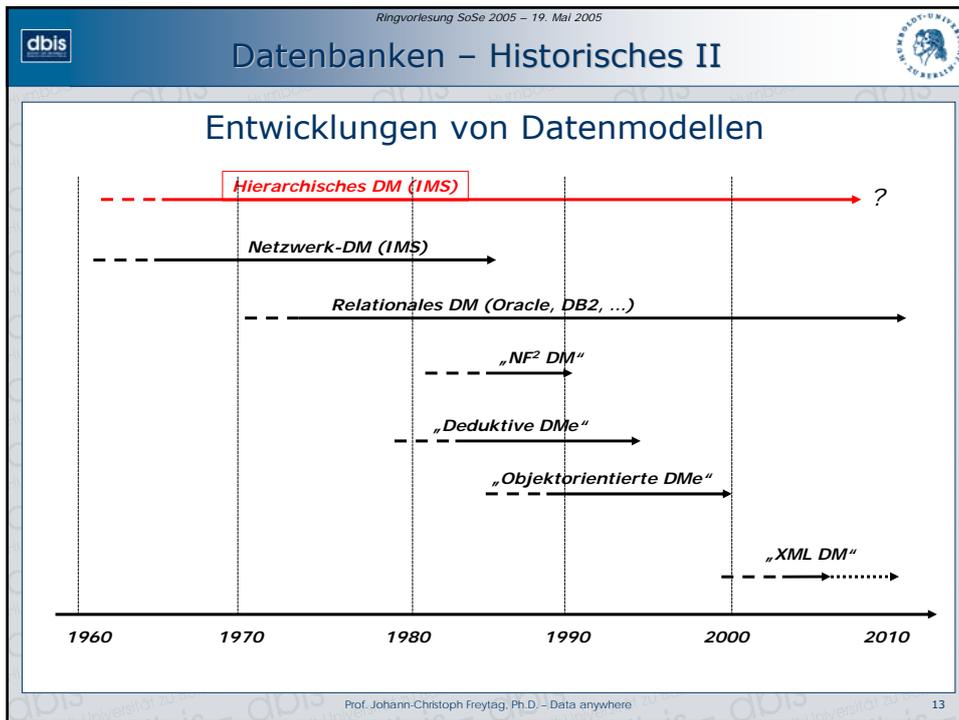
IMS (IBM)



CODASYL-DB

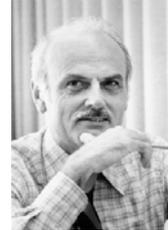


- Hierarchisch – Netzwerk: Beides „Navigierende Modelle“
 - Programmierer muss Hierarchie/Netzwerk kennen
 - Komplexe Sprache mit vielen Feinheiten (Effizienz)



- Relationales Modell
Edgar F. "Ted" Codd (1923 - 2003)

- Britischer Mathematiker
- 1949 – 1979 IBM Mitarbeiter
- Ab 1980 Codd&Date Consulting
- 1981 ACM Turing Award
- IBM Fellow, ...



- Wie alles anfing ...

A relational model of data for large shared data banks

[Communications of the ACM, Volume 13, Issue 6 \(June 1970\), Pages: 377 - 387](#)

It provides a means of describing data with **its natural structure only**--that is, **without superimposing any additional structure for machine representation purposes**. Accordingly, it provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation on the other.

- Grundvision des Relationalen Modells*
(Nach der Veröffentlichung bzw. ACM Turing Award Rede)

- **Alle** Informationen können als Werte in Relationen (Tabellen) dargestellt werden
- **Keine** Information soll durch Zeiger (pointer), Indexe, Links oder durch Ordnen von Objekten repräsentiert werden.
- "Zugriffsmethoden" sollen ausschließlich zur Verbesserung der Performanz genutzt werden, sie dürfen aber keine essentielle Information enthalten.

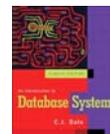
- Das Relationale Modell
 - „Tabellenmodell“

supply	(supplier	part	project	quantity)
1	2	5	17	
1	3	5	23	
2	3	7	9	
2	7	5	4	
4	1	1	12	

FIG. 1. A relation of degree 4

- Mit deklarativer Anfragesprache: **Was – Nicht Wie!!**
 - Tupel-/Domänenkalkül & Relationale Algebra
 - SQL kam erst später (1974 in Form von SEQUEL)
- Grundlagen in Theorie und Praxis
 - Hat Theoretiker und Systemimplentierer gleichermaßen fasziniert und zur Forschung angeregt!!

- **Wie alles anfing ... schwer...**
 - IBM: kein Interesse – IMS war das strategische Produkt
 - San Jose war weit weg von der Ostküste (head quarters)
 - Netzwerk-Datenbanken hatten „Hochkonjunktur“
 - Kein existierendes System als Beweis, das es „funktioniert“
- Codd's Strategie
 - Veröffentlichungen:
 - mehr als ein Paper
 - Wichtigster Partner: Chris Date
 - Rededuelle mit Charles Bachmann
 - Netzwerk vs. Relational
- ... und was noch half: Systeme
 - Entwicklung von System R ab 1975 ([link](#)) in San Jose, CA
 - Entwicklung von INGRES (M. Stonebraker, UC Berkeley)



- System R

- Entwurf (Design) einer DBMS Architektur
 - ... wie sie bis heute immer noch existiert
 - Anfrageoptimierung
 - Transaktionsverwaltung
- Effiziente Realisierung
 - Wichtiger Beitrag: B-Bäume (B = Balanciert)
 - Entwickelt durch Rudolf (Rudi) Bayer ([link](#)) & Edward M. McCreight



Organization and Maintenance of Large Ordered Indices, Acta Inf. 1: 173-189 (1972)

- Wichtigsten Mitarbeiter des Projektes System R



Don Chamberlin



Pat Selinger



Jim Gray



Raimond Lorie

- DB Entwicklungen der IBM
- Erstes DBMS-Produkt: SQL/DS auf VM (1981)
 - Halbherzig und wenig performant
- Weiteres Produkt: DB2 - 1983
 - Strategische Plattform MVS (Großrechner)
- Forschungsprojekte
 - **System R*** : Verteiltes RDBMS (ab 1979) auf MVS
 - Der Zeit (zu) weit voraus
 - Verteilte Transaktionsverarbeitung/Anfragebearbeitung
 - **Starburst** (ab ca. 1983)
 - Entwicklung auf PC/Workstation
 - Implementiert in C-Nutzung von TCP/IP
 - Ziel: Erweiterbares Hochleistungs-DBMS
 - ADTs (abstract data types), UDFs (user-defined functions)
 - Produkt auf AIX & anderen UNIX/NT-Plattformen (DB2/UDB)

Relationale Technologie VIII

INGRES (UC Berkeley)

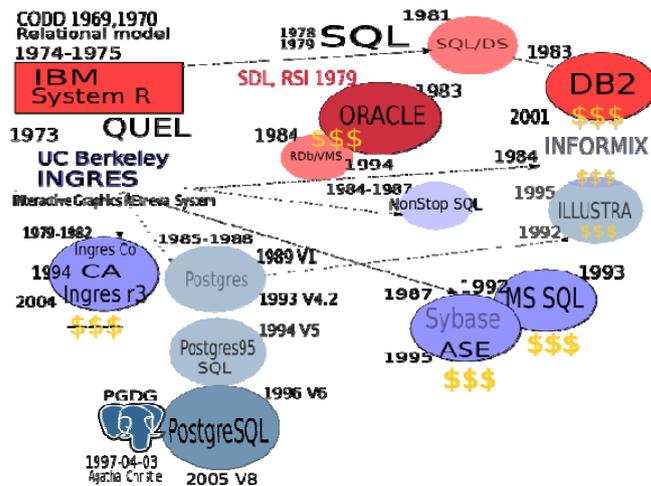


- Leiter: Mike Stonebraker
 - Professor UC Berkeley (ehem.)
 - Immer noch aktiv in Forschung (MIT) und Wirtschaft
- Viele Ph.D.Arbeiten, u.a.
 - Bob Epstein (Gründer Sybase)
- INGRES Forschungsprojekt
 - Große Konkurrenz zu IBM
 - Entwickelte alle Komponenten eines DBMSs
- Gründung der Firma INGRES
 - Kommerzielle weniger erfolgreich
 - Aufgekauft von Computer Associates (CA) (ca. 1995)
- Entwicklung von Postgres
 - Firma Illustra – an Informix verkauft



Relationale Technologie IX

- Systementwicklungen



- Oracle

- Gegründet von Larry Ellison (1979)
 - Zweites (erstes??) relationales Datenbankprodukt
 - Gründung als Garagenfirma
 - Zunächst: Software Development Laboratories (SDL)
 - 1979: Relational Software, Inc. (RSI)
 - 1983: Oracle (früherer Codename des Projektes)
- Auf Grund der Veröffentlichungen der IBM Forschung
- Strategie: DBMS „auf allen Plattformen“ (portability)
 - In C implementiert (damals revolutionär)
 - Erste HW-Plattform: DEC PDP-11 – UNIX
- Einer der außergewöhnlichsten Geschäftsleute der USA



Deutsche Entwicklungen
Meist an Universitätsentwicklungen & Ausgründung

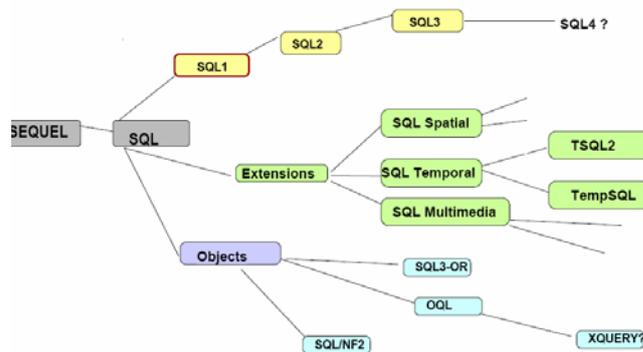
- SAPDB (Entwicklung in Berlin)
 - Entwickelt durch Rudolf Munz an der TU Berlin
 - Verteiltes DBMS (WELL System)
 - Firmengründung ca. 1979
 - Teil von Nixdorf, Siemens, Software AG und SAP
 - Heute als MaxDB™ durch MySQL vertrieben
- TransAction Software GmbH, München
 - 1987 - Gründung durch ehemalige Doktoranden (& Prof. R. Bayer)
 - Unternehmen mit einem Hochleistungs-RDBMS
 - Weiterentwicklung durch R. Bayer, TU München
- PASCAL-R
 - Entwickelt von Prof. Joachim Schmidt, Univ. Hamburg ab ca. 1977
 - Erweiterung der und Einbettung in die Programmiersprache PASCAL
 - DB-Zugriff durch prädikatenlogische Ausdrücke
 - Elegante Lösung ohne „Zwei-Welten-Phänomen“

- **Relationale Technologie und theoretische Arbeiten**

- Codd: Mathematiker
 - Relationales Modell: theoretisch fundiert
- Ermöglichte vielfältige theoretische Arbeiten
 - Datenbank(Schema)-Entwurf
 - Abhängigkeitstheorie
 - Funktionale Abhängigkeiten (FDs)
 - Multi-Value Dependencies (MVDs), ...
 - Normalformen
 - Anfragesprachen
 - Mächtigkeit
 - Weiterentwicklungen (deduktive Sprachen)
- Bekanntester Vertreter: Jeffrey D. Ullman
 - Viele Artikel, mehrere Bücher



SQL-Entwicklungen



Ringvorlesung SoSe 2005 – 19. Mai 2005

dbis 

Agenda

- Überblick
- Plattentechnologie
- Datenmodelle & Datenbanken
- Relationales Technologie
- Weitere Entwicklung

Prof. Johann-Christoph Freytag, Ph.D. – Data anywhere 27

Ringvorlesung SoSe 2005 – 19. Mai 2005

dbis 

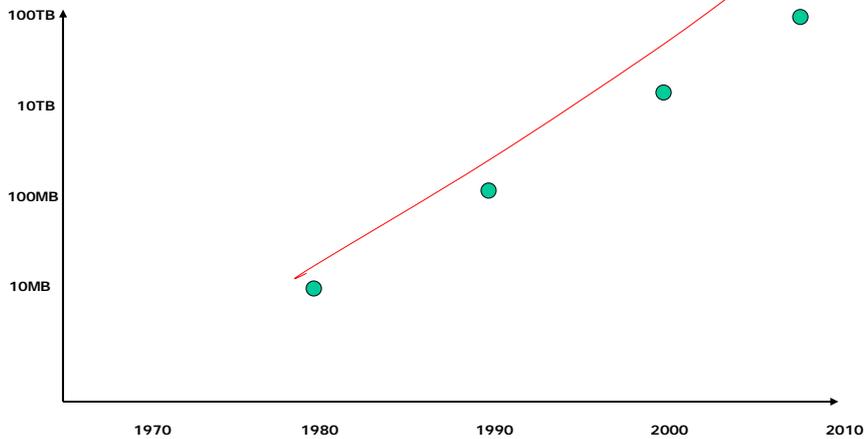
Relationale Entwicklung I

- Beobachtung:
 - Relationale Sprachen und die damit verbundene Technologie realisieren
 - maschinenunabhängige, **deklarative** Programmierung verbunden
 - mit Hochleistungsansprüchen, die erfüllt werden
 - Dieser Ansatz erlaubt kontinuierliche Anpassung an neue „Realisierungstechnologien“
- Kommerzielle Nutzung weltweit in allen Branchen
 - Informationsintegration, Web, Grid, P2P
- Ständiger Schub an neuen Innovationen
- Enge Verzahnung von „Theorie und Praxis“
 - Ständiger Austausch zwischen Universitäten und Firmen

Prof. Johann-Christoph Freytag, Ph.D. – Data anywhere 28



Größe von Datenbanken



[Let's have a look at the Web](#)



Beispiele für große Datensammlungen

- Jim Gray & SkyServer ([link](#))
 - 40TB an Daten auf Microsoft SQLServer
- Jim Gray & TerraServer ([link](#))

Ringvorlesung SoSe 2005 – 19. Mai 2005

dbis

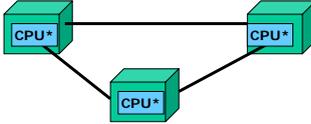


Relationale Entwicklung IV

- Weiterentwicklungen
 - Parallele DBMS (1990-2000)
 - Shared memory: bis zu 32 Prozessoren



- Shared nothing (Cluster): beliebig



- Neuere Entwicklungen (ab 2000)
 - XML als Datenbanksprache: Ablösung von SQL??
 - Daten in P2P-Umgebungen
 - Informationsintegration

Prof. Johann-Christoph Freytag, Ph.D. – Data anywhere 31

Ringvorlesung SoSe 2005 – 19. Mai 2005

dbis



Caveats

- Trotz aller erfolgreichen RDBMS-Entwicklungen
 - Relationale DBMS speichern nur ca. 5-10 % aller Daten weltweit (geschätzt)
 - Vorherrschend: dateibasiert - Warum??
- RDBMS – ein „commodity item“?
 - Meine These:
 - „Jein“ – immer noch viele offene Forschungsfragen, die einen Einfluß auf die Qualität des Produktes haben werden
 - „Beweis“: MS und IBM investieren immer noch in Forschungskapazitäten im Bereich DB
- Neue Herausforderungen
 - Große (!!) Datenmengen & komplexe Anfragen
 - Security & Privacy
 - Daten- & Informationsintegration

Prof. Johann-Christoph Freytag, Ph.D. – Data anywhere 32

The end ...

Fragen ?

