Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Beauty is our Business

Organisatorisches und Einführung

22.10.2009

Felix Naumann

Das Motto

... Wenn wir uns klarmachen, dass der Kampf gegen Chaos, Durcheinander, und unbeherrschte Kompliziertheit eine der größten Herausforderungen der Informatik ist, müssen wir zugestehen:

„Beauty is our Business“.

Edsger W. Dijkstra, 1978

Auch:

“Computer Science is no more about computers than astronomy is about telescopes.”

Motivation

- Vorbereitung auf das Master Studium
  - Wissenschaftlich Arbeiten
- Vorbereitung auf den Beruf
  - Ideen „verkaufen“
- Interesse an den Themen

Überblick

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Kurzvorstellung der Themen

Forschungsgruppe Informationsintegration

Felix Naumann
- naumann@hpi.uni-potsdam.de
- Raum A-1.13 (über Frau Pamperin)
- HU, IBM Almaden, DFG/JP/HU, HPI

Universität Potsdam, Hasso-Plattner-Institut
- Fachgebiet „Informationssysteme“

Forschungsthemen (gleich mehr):
- Datenqualität
- Informationsintegration
- Peer Data Management
- Search
- http://www.hpi.uni-potsdam.de/~naumann/

Information Systems Team

Paul **Führing**

project **ViQTOR**

DQ Annotation & Assessment

Katrin **Heinrich**

Prof. Felix **Naumann**

Information Integration

Data Fusion

Jens **Bleiholder**

project **fusem**

Dustin Lange

Schufa

Johannes Lorey

project **HumMer**

Armin **Roth**

project **System P**

Peer Data Management Systems

Information Quality

Data Profiling

IBM

Data Profiling & Cleaning

Christoph Böhm

Matching

Data Integration for Life Science Data Sources

project **PoSR**

Service-Oriented Systems

Tobias **Vogel**

Forschungskolleg

Mohammed **AbuJarour**

Ontologies, Profiling

Frank **Kaufer**

project **Aladin**

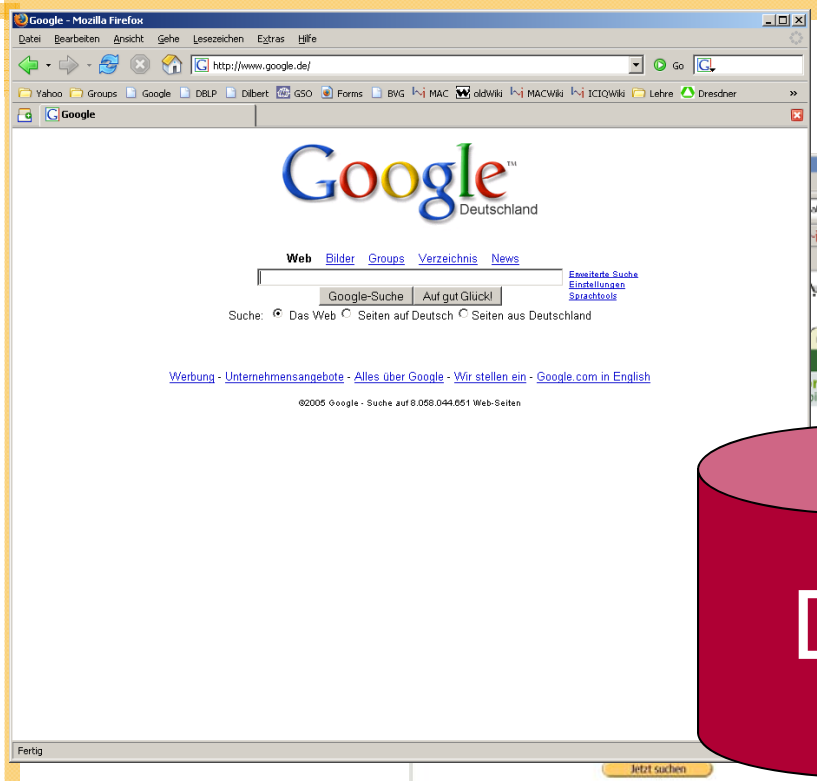
Jana **Bauckmann**

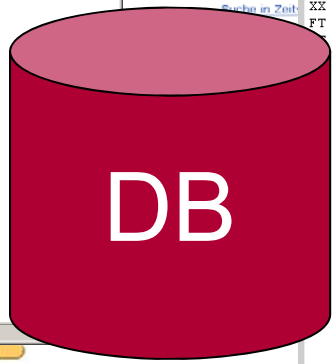
Alexander **Albrecht**

ETL Management

Data Profiling for Schema Management

Was sind Informationssysteme?



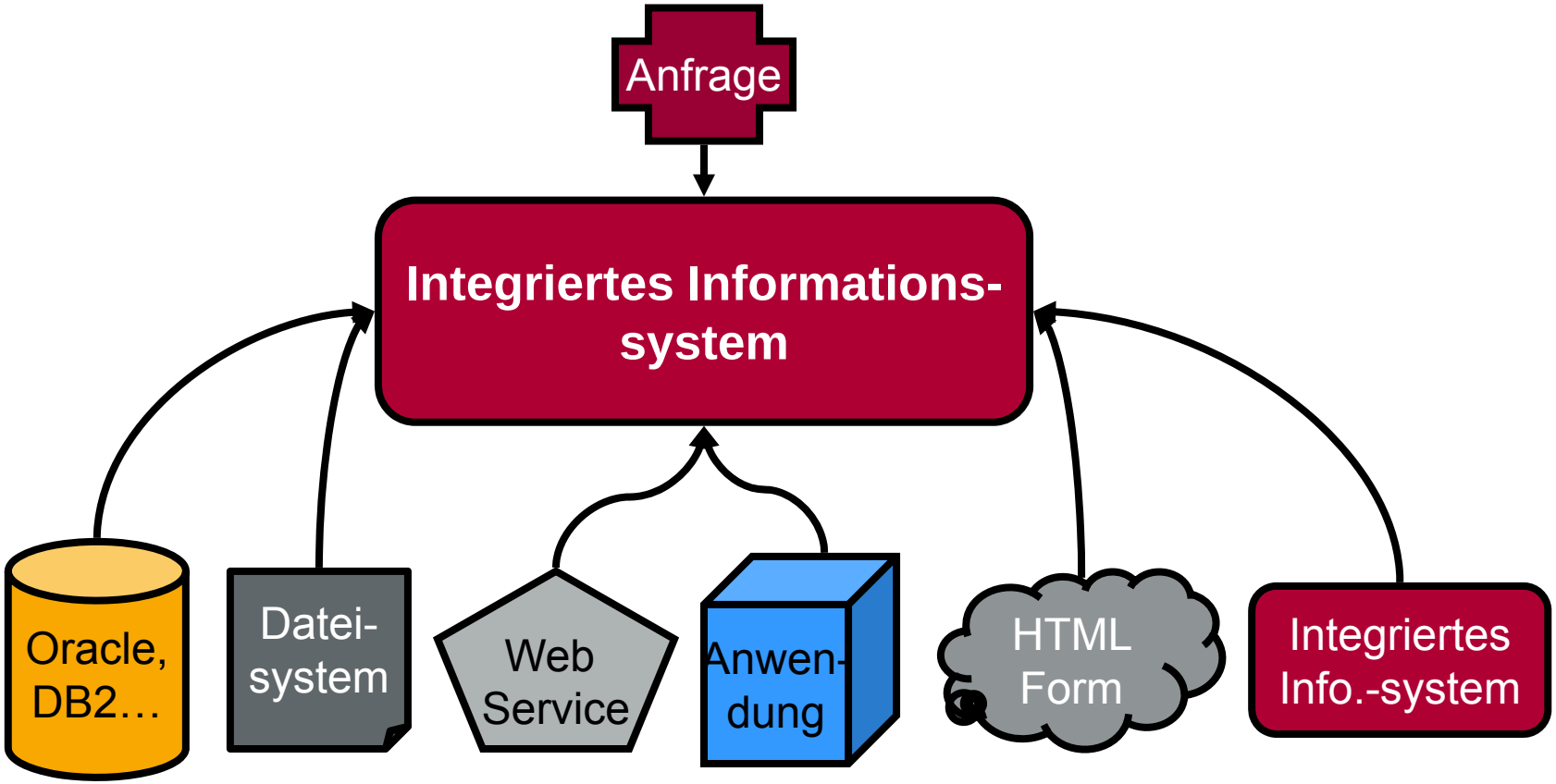



```
ID   RNGTPCHI   standard; RNA; ROD; 1016 BP.
XX
DT   01-AUG-1991 (Rel. 28, Created)
DT   04-MAR-2000 (Rel. 63, Last updated, Version 2)
XX
DE   Rat GTP cyclohydrolase I mRNA, complete cds.
XX
KW   GTP cyclohydrolase I.
XX
OS   Rattus norvegicus (Norway rat)
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia;
OC   Eutheria; Rodentia; Sciurognathi; Muridae; Murinae; Rattus.
XX
RN   [1]
RP   1-1016
RX   MEDLINE; 91093270.
RX   PUBMED; 1985963.
RA   Hatakeyama K., Inoue Y., Harada T., Kagamiyama H.;
RT   "Cloning and sequencing of cDNA encoding rat GTP cyclohydrolase I: The
RT   first enzyme of the tetrahydrobiopterin biosynthetic pathway";
RL   J. Biol. Chem. 266(2):765-769(1991).
XX
FT   CDS             128..853
FT                   /codon_start=1
FT                   /db_xref="GOA:P22288"
FT                   /db_xref="SWISS-PROT:P22288"
FT                   /EC_number="3.5.4.16"
FT                   /gene="GTP cyclohydrolase I"
FT                   /product="GTP cyclohydrolase I"
FT                   /protein_id="AAA41299.1"
FT                   /translation="MEKPRGVRCTNGFPERELPRPGASRPAEKSRPPEAKGAQPADAWK
FT                   AGRPRSEEDNELNLPNLAAAYSSILRSLGEDPQRQGLLKTPWRAATAMQFFTKGYQETI
FT                   SDVLNDAIFDEDHDEMVIVKDIDMFSMCEHHLVPFVGRVHIGYLPNKQVLGLSKLARIV
FT                   EIYSRRLQVQERLTKQIAVAITEALQPAGVGVVIEATHMCMVMRGVQKMNSKTVTSTML
FT                   GVFREDPKTREEFLTLIRS"
Sequence 1016 BP; 236 A; 279 C; 291 G; 210 T; 0 other;
gacttcgaac ctcattcggt gcagaactcc tgtcccggtg acagccacag gtcacggccg        60
ccggctaagc cgagccgcag cgcttggtag caccttaggg tgtctcggga gcaatcgcgc       120
cgggtccatg gagaagccgc ggggtgtaag gtgcaccaat gggttccccg agcgggagct       180
   ...
catcaggagc tgaacttccg tgtgcgagcc ccggtttgca gacccccgct gaggccagcg       900
ttatctgtct cgattgtaca ttccagttcc agttggtata cttgtcaact ttatttctca       960
ccatgaattg tatttaataa ttatttatag agatgtcaaa taaaggtgat caacttt         1016
//
```

Molecule type
Name
Date of creation and last update
Free text description
Keywords describing the molecule
Organism
Article the sequence was published in
Structural annotation (coding sequence)
Link to functional annotation of resulting protein
Translated protein sequence
Sequence of bases

```
<buch>
        <isbn>0-201-318051</isbn>
        <titel>XML und Datenbanken</titel>
        <autor>Klettke/Meyer</autor>
</buch>
```

Integrierte Informationssysteme



Schematische und Daten-Heterogenität

Variante 1

Männer

Vorname	Nachname
Felix	Naumann
Jens	Bleiholder

Frauen

Vorname	Nachname
Melanie	Weis
Jana	Bauckmann

Variante 2

Personen

Vorname	Nachname	Männl.	Weibl.
Felix	Naumann	Ja	Nein
Jens	Bleiholder	Ja	Nein
Melanie	Weis	Nein	Ja
Jana	Bauckmann	Nein	Ja

Variante 3

Personen

Vorname	Nachname	Geschlecht
Felix	Naumann	Männlich
Jens	Bleiholder	Männlich
Melanie	Weis	Weiblich
Jana	Bauckmann	Weiblich

Schematische und Daten-Heterogenität

Variante 1

Männer

Vorname	Nachname
Felix	Naumann
Jens	Bleiholder

Frauen

Vorname	Nachname
Melanie	Weis
Jana	Bauckmann

Variante 2

Personen

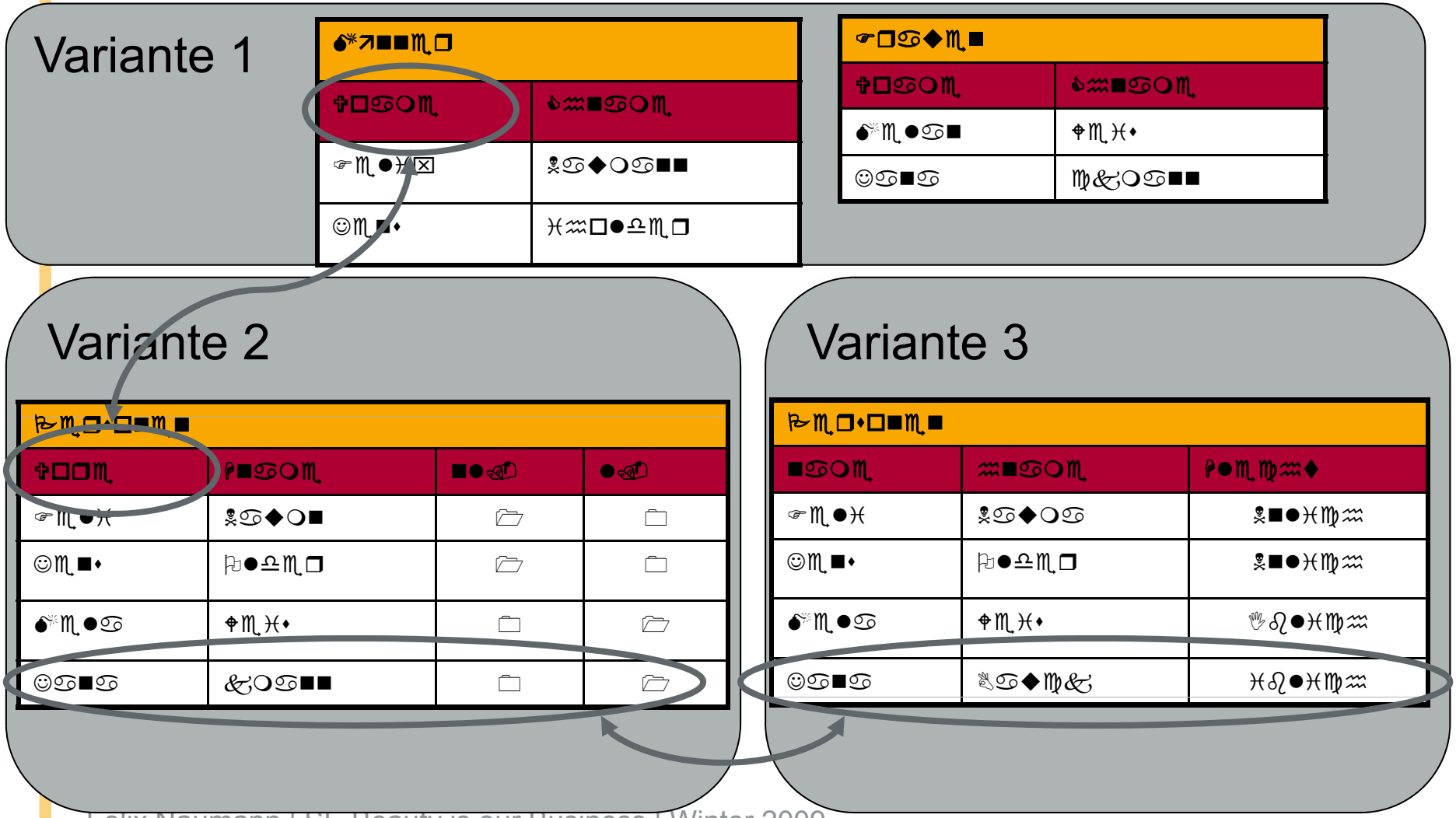
FirstNa	Name	male	femal
Felix	Naumann	Ja	Nein
Jnes	Bleiho.	Ja	Nein
Melanie	Weiß	Nein	Ja
Jana	baukman	Nein	Ja

Variante 3

Personen

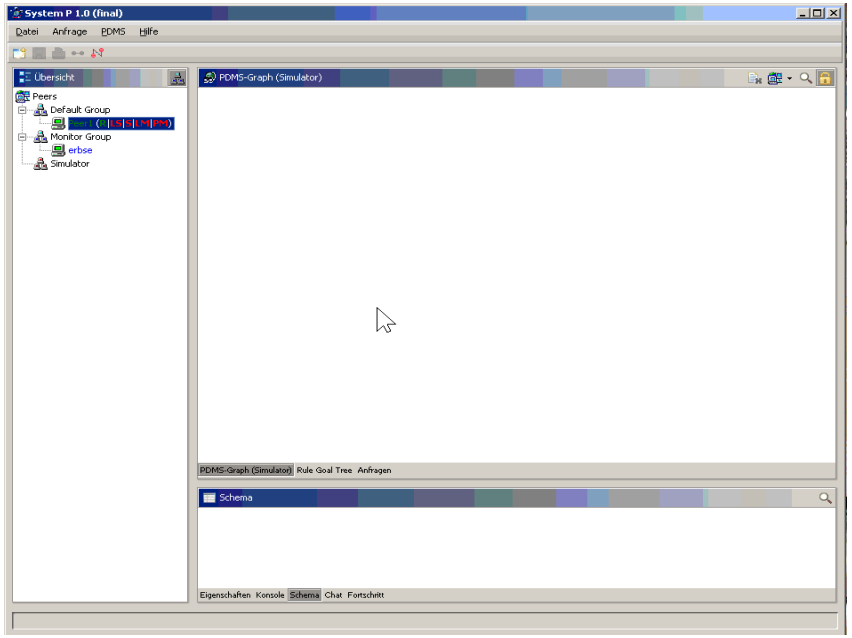
VN	NN	SEX
F.	Naumann	Männlich
J.	Bleiholder	Männlich
M.	Weis	Weiblich
J.	Bauckmann	Weiblich

Schematische und Daten-Heterogenität

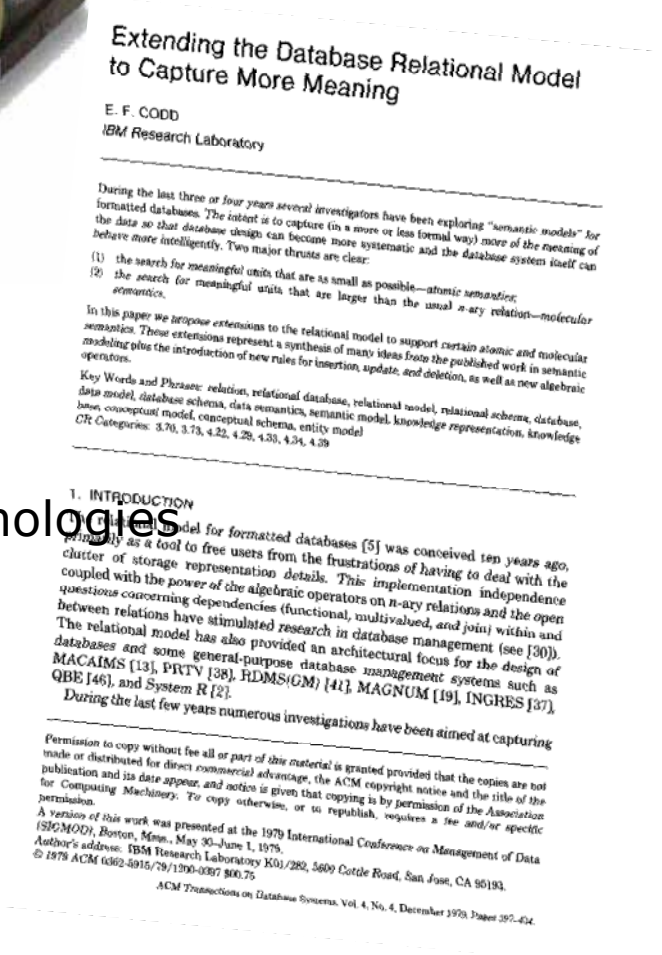


Weitere Forschungsthemen

- Informationsintegration
  - Schema Matching
  - Duplikaterkennung
  - Datenfusion
- Datenqualität
- Peer Data Management
- Life Sciences: Aladin
- Search



Lehrveranstaltungen in diesem Semester



Extending the Database Relational Model to Capture More Meaning

E. F. CODD

IBM Research Laboratory

During the last three or four years several investigators have been exploring "semantic models" for formatted databases. The intent is to capture (in a more or less formal way) more of the meaning of the data so that database design can become more systematic and the database system itself can behave more intelligently. Two major thrusts are clear:

(1) the search for meaningful units that are as small as possible—atomic semantics;
(2) the search for meaningful units that are larger than the usual n-ary relation—molecular semantics.

In this paper we propose extensions to the relational model to support certain atomic and molecular semantics. These extensions represent a synthesis of many ideas from the published work in semantic modeling plus the introduction of new rules for insertion, update, and deletion, as well as new algebraic operators.

Key Words and Phrases: relation, relational database, relational model, relational schema, database, data model, database schema, data semantics, semantic model, knowledge representation, knowledge base, conceptual model, conceptual schema, entity model

CR Categories: 3.70, 3.73, 4.22, 4.29, 4.33, 4.34, 4.39

1. INTRODUCTION

The relational model for formatted databases [5] was conceived ten years ago, primarily as a tool to free users from the frustrations of having to deal with the clutter of storage representation details. This implementation independence coupled with the power of the algebraic operators on n-ary relations and the open questions concerning dependencies (functional, multivalued, and join) within and between relations have stimulated research in database management (see [30]). The relational model has also provided an architectural focus for the design of databases and some general-purpose database management systems such as MACAIMS [13], PRTV [38], RDMS(GM) [41], MAGNUM [19], INGRES [37], QBE [46], and System R [2].

During the last few years numerous investigations have been aimed at capturing

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.

A version of this work was presented at the 1979 International Conference on Management of Data (SIGMOD), Boston, Mass., May 30–June 1, 1979.

Author's address: IBM Research Laboratory K01/282, 5600 Cottle Road, San Jose, CA 95193.

© 1979 ACM 0362-5915/79/1200-0397 $00.75

ACM Transactions on Database Systems, Vol. 4, No. 4, December 1979, Pages 397–434.

- Vorlesungen
  - DBS II
- Seminare
  - Bachelor: Beauty is our Business
  - Master: Advanced Map/Reduce algorithms on Hadoop
  - Master: Emerging Web Service Technologies
  - Forschungsseminar / InfoLunch
- Bachelorprojekte
  - ETL Management mit BMW
  - Extreme Data Integration mit IBM

Überblick

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Kurzvorstellung der Themen

Vorstellung Teilnehmer

- Sebastian Blessing
- Lena Herscheid
- Thomas Stoff
- Stefan George
- Felix Leupold
- Christian Wegner

Seminarleistungen

Lesen
- Paper lesen und verstehen
- Verwandte Literatur lesen und verstehen
- Mindestens eine individuelle Besprechung mit mir
- Aktive Teilnahme an anderen Vorträgen

1/5 Note

Vortragen
- Kurzvorstellung der Literatur (je 2 min)
- Mindestens eine Folien-Besprechung mit mir
  - Spätestens 1 Woche vor Vortrag
- 30 min. Vortrag am jeweiligen Termin
  - + 15 min Diskussion

2/5 Note

Ausarbeitung
- Gliederungs-Besprechung
  - Spätestens 2 Wochen vor Abgabetermin
- 8-10-seitige Ausarbeitung bis zum 12.3.2010
- Unter Verwendung der LaTeX-Vorlage im WWW (neu)

2/5 Note

Aktive Teilnahme an sämtlichen gemeinsamen Terminen

Feedback

Fragen bitte jederzeit!

- Während des Seminars
- Während individueller Besprechungen
  - Termin bei Katrin Heinrich ausmachen
    - office-naumann@hpi.uni-potsdam.de
    - 0331 / 5509 280
  - Muss nicht zur Sprechstunde sein!
- Sprechstunde
  - Dienstags 15:00 – 16:00
  - Raum A-1.13
  - Am liebsten mit Anmeldung
- Email: naumann@hpi.uni-potsdam.de

Überblick

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Kurzvorstellung der Themen

Wissenschaftliche Texte lesen

- Fachartikel
  - Journale, Konferenzen und Workshops
  - Entstehung:
    - Forschungsvorhaben
    - Begutachtungsprozess
- Struktur eines Artikels
  - Kritisches Lesen
  - Experimente
- Literatur
  - (Online-) Recherche
- Englisch

Wissenschaftlichen Vortrag halten

- Gliederung
  - Die Kunst des Weglassens
- Foliengestaltung
  - Powerpoint, Keynote, usw.
- Zeit einhalten
- Techniken zur Vorbereitung
- Techniken während des Vortrags

Ausarbeitung schreiben

- Gliederung
- Schreibstil
  - Objektives
  - Subjektives
- Plagiate
- LaTeX
  - denn „Beauty is our Business“

Zeitlicher Ablauf – Termine

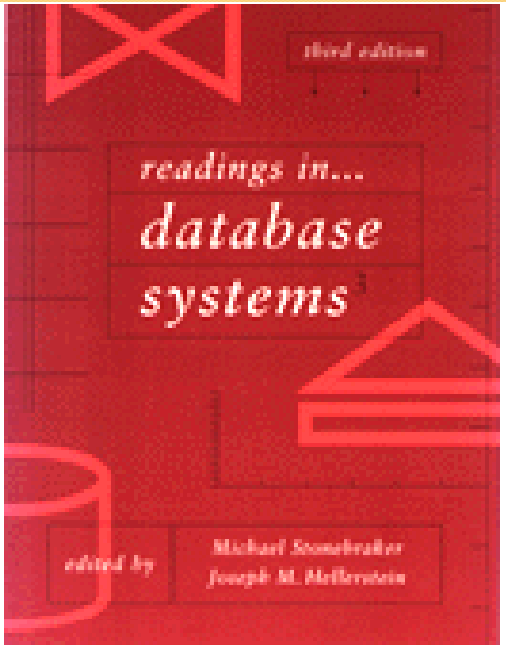
- Part 1: Lesen
  - Termin zur Literaturkritik
  - *Elevatorpitch*
- Part 2: Vortragen (Zwei pro Termin)
  - Jeweils 30 Minuten Vortrag + Diskussion
- Part 3: Schreiben
  - Vorstellung der Gliederung
  - Ausarbeitung / Seminararbeit
- Vortragsreihenfolge
  - Siehe Web-Page
- Dazwischen Vorträge von mir
  - Lesen wissenschaftlicher Texte
  - Vortragstechniken
  - LaTeX (bei Bedarf)
  - Tipps zur Gliederung und Ausarbeitung

Überblick

- Vorstellung der Arbeitsgruppe
- Organisatorisches
- Ziele des Seminars
- Kurzvorstellung der Themen

Reading in Database Systems

- Editors
  - Joseph M. Hellerstein
  - Michael Stonebraker
- http://redbook.cs.berkeley.edu/
- Zusammenstellung wichtigster Forschungsarbeiten
- Dritte Auflage

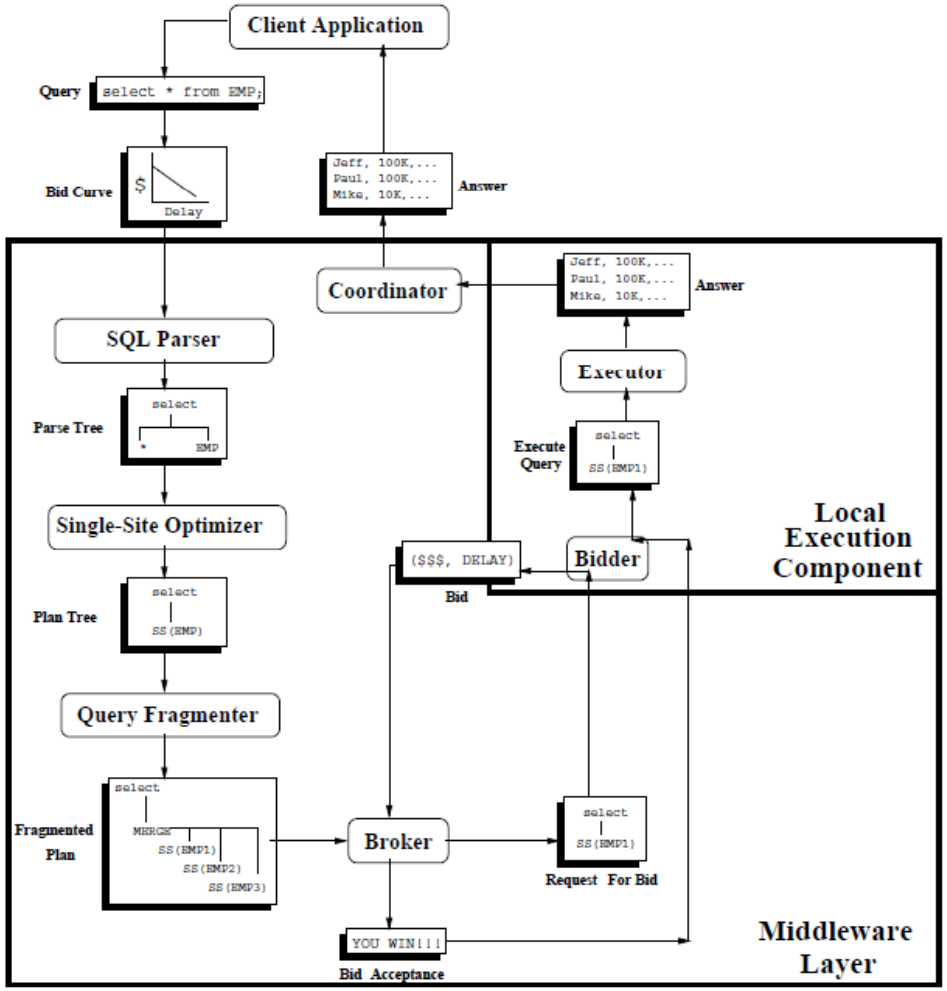


Meine Auswahl

1. Mariposa: Neue Architektur
2. R*-Trees: Mehrdimensionale Indizes
3. The Five-Minute Rule: Datenbankoptimierung
4. BIRCH-clustering
5. The Apriori Algorithm: Data Mining
6. Eddies: Dynamische Optimierung

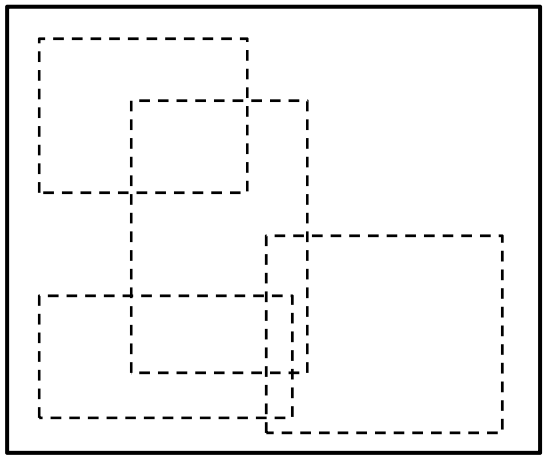
Mariposa

- Michael Stonebraker
- Neuartige Architektur für ein weit verteiltes Informationssystem
- Mikroökonomisches Prinzip zur Anfrageoptimierung
  - Datenhaltung
  - Anfragebearbeitung



R*-Trees

- Hans-Peter Kriegel
- An Efficient and Robust Access Method for Points and Rectangles
- Klasischer R-Baum unterteilt 2- oder mehrdimensionale Räume in Regionen
  - Jede Form, meist Rechteck
  - Regionen und Subregionen
  - Subregionen decken nicht alles ab
  - Subregionen können sich überlappen
- Operationen: Anfragen und Updates



- Bisher (R-Baum): Minimiere Fläche der Rechtecke
- Nun (R*-Baum): Neue Kriterien, wie Umfang, Überlappung, usw.

The Five-Minute-Rule

THE 5 MINUTE RULE FOR TRADING MEMORY FOR DISC ACCESSES
and
THE 10 BYTE RULE FOR TRADING MEMORY FOR CPU TIME

Jim Gray
Franco Putzolu
Tandem Computers, Cupertino, CA, USA

- Jim Gray & Goetz Graefe
- The Five-Minute Rule Ten Years Later, and Other Computer Storage Rules of Thumb
- Optimale Verweilzeit einer Seite im Cache
  - Tradeoff: Preis und Bandbreite von RAM und Disk
- 1987: „Pages referenced every five minutes should have been kept in memory rather than reading them from disk each time”
- 1997: “In summary, the five-minute rule still seems to apply to randomly accessed pages, primarily because page sizes have grown from 1KB to 8KB to compensate for changing technology ratios.“
- 2007: „Flash memory falls between traditional RAM and persistent mass storage based on rotating disks in terms of acquisition cost, access latency, transfer bandwidth, spatial density, power consumption, and cooling costs. “

BIRCH Clustering

- Raghu Ramakrishnan
- Ziel: Dichtbesetzte Gebiete in mehrdimensionalen Räumen finden.
- BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies
- Features:
  - Good clustering in first scan, each additional scan improves quality
  - Scales gracefully to reside in main memory
  - Outlier handling
  - Incremental clustering

Apriori: Data Mining

Fast Algorithms for Mining Association Rules

- Rakesh Agrawal
- Bahnbrechendes Papier
- Setzte intensive Forschung zu Data Mining in Gang
- Zwei Algorithmen zur schnellen Endeckung von Assoziationsregeln
  - Apriori
  - AprioriTid

Eddies – Dynamic Optimization

- Joe Hellerstein
- Eddies: Continuously Adaptive Query Processing
- Eigenschaften großer, verteilter Systeme ändern sich noch während der Anfragebearbeitung
- Joinreihenfolgen dynamisch verändern
  - Moments of symmetry

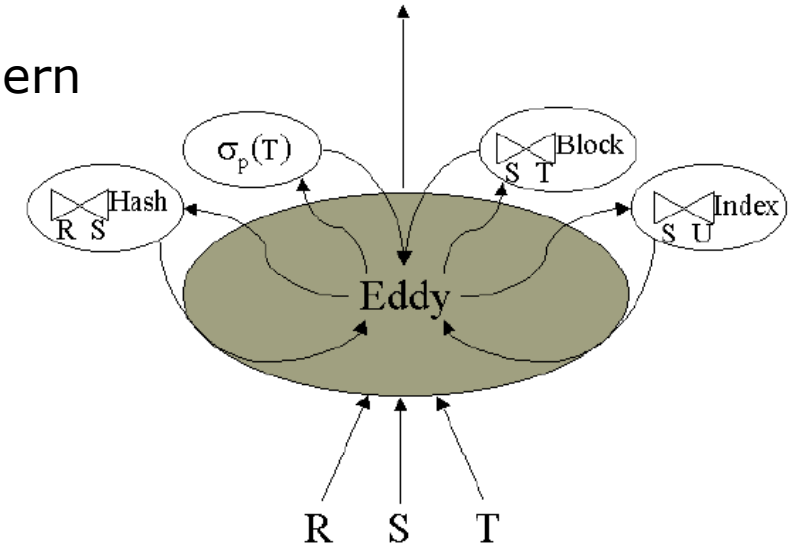


Figure 1: An eddy in a pipeline. Data flows into the eddy from input relations R, S and T. The eddy routes tuples to operators; the operators run as independent threads, returning tuples to the eddy. The eddy sends a tuple to the output only when it has been handled by all the operators. The eddy adaptively chooses an order to route each tuple through the operators.

Nächste Schritte

- Sie: E-Mail an mich
  - Mit Wunschliste (Platz 1 – 3)
    - Daraus: Themenvergabe
  - Bis Mitternacht
- Ich: Themenvergabe bald
  - E-Mail Benachrichtigung
  - http://www.informatik.uni-trier.de/~ley/db/index.html
- Ich: Vortrag zum „Lesen wissenschaftlicher Texte“