Specialization Area
Information Systems

Databases and Information Systems Lab
(Prof. Härder & Prof. Michel)

Heterogeneous Information Systems Lab
(Prof. Deßloch)
What are Information Systems?
Our Foundations Course: (Advanced) Database Systems

Concurrency Control, Recovery, and Buffer Management

\[ T_1 \quad T_2 \quad \ldots \quad T_n \]

Transaction Manager (TM)

Scheduler

Data Manager (DM)

Recovery Manager

Buffer Manager

Database

Query Rewriting/Unnesting/Optimization

\[
\begin{align*}
\text{SELECT DISTINCT} & \quad \text{C.cname, (SELECT count(*)}) \\
\text{FROM} & \quad \text{Product P} \\
\text{WHERE} & \quad \text{P.cid = C.cid)}
\end{align*}
\]

FROM Company C

Probabilistic Databases

<table>
<thead>
<tr>
<th>Name</th>
<th>Bird</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mary</td>
<td>Bird-1</td>
<td>{Finch: 0.8, Toucan: 0.2}</td>
</tr>
<tr>
<td>Susan</td>
<td>Bird-2</td>
<td>{Nightingale: 0.65, Toucan: 0.35}</td>
</tr>
<tr>
<td>Paul</td>
<td>Bird-3</td>
<td>{Kookaburra: 0.55, Toucan: 0.45}</td>
</tr>
</tbody>
</table>

Multidimensional and Metric Data Structures

Join Order Optimization

<table>
<thead>
<tr>
<th>Join Order</th>
<th>( C_{out} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( R_1 \times R_2 )</td>
<td>2</td>
</tr>
<tr>
<td>( R_2 \times R_3 )</td>
<td>200</td>
</tr>
<tr>
<td>( R_3 \times R_4 )</td>
<td>2</td>
</tr>
<tr>
<td>( ((R_1 \times R_2) \times R_3) \times R_4 )</td>
<td>24</td>
</tr>
<tr>
<td>( ((R_2 \times R_3) \times R_1) \times R_4 )</td>
<td>222</td>
</tr>
<tr>
<td>( (R_1 \times R_2) \times (R_3 \times R_4) )</td>
<td>6</td>
</tr>
</tbody>
</table>
The Big Data Era: A few Numbers ...

Data Volume

• Google: 15 000 PB (=15 Exabytes)
• Facebook: 300 PB
• Ebay: 90 PB
• Spotify: 10 PB

Data Processed per Day

• Google: 100 PB
• Ebay: 100 PB
• NSA: 29 PB
• Facebook: 600 TB
• Twitter: 100 TB
• Spotify: 2,2 TB

MB = $10^6$ Byte
GB = $10^9$ Byte
TB (Terabyte) = $10^{12}$ Bytes
PB (Petabyte) = $10^{15}$ Bytes
EB (Exabyte) = $10^{18}$ Bytes
Example: Critical Data Volume

• Assume we have **10 TB data stored on our hard disk**.

• **Now we want to analyze this data!**

• Using a hard disk having **100MB/s read rate, solely reading takes**
  • 100000 seconds, or
  • 1666 minutes, or
  • 27 hours
Google in 1998 vs. Today

http://flickr.com/photos/jurvetson/157722937/

http://www.google.com/about/datacenters/inside/index.html
Assume a disk fails once per year and we have $n$ disks. What is the probability that at least one fails today?

- $n=1$: 0.0027
- $N=100$: 0.239
- $N=1000$: 0.9356
- $N=10000$: ~ 1.0
Course Distributed Data Management (NoSQL / Cloud / Big Data)

**Map** \((k1,v1) \rightarrow \text{list}(k2,v2)\)

**Reduce** \((k2, \text{list}(v2)) \rightarrow \text{list}(k3, v3)\)

**MapReduce**

**Spark**

**Data Placement via Consistent Hashing**

**Time Stamps and Vector Clocks**

**CAP Theorem**

**WARS Model and Consistency Levels**

**Distributed Consensus (PAXOS)**

(Lease Lamport)
Course Information Retrieval and Data Mining

Relevance Assessment

PageRank

Decision Trees

Vector Space Model

\[ \text{sim}(q, d) = \frac{q \cdot d}{\|q\| \|d\|} = \frac{\sum_{i=1}^{\|V\|} q_i d_i}{\sqrt{\sum_{i=1}^{\|V\|} q_i^2} \sqrt{\sum_{i=1}^{\|V\|} d_i^2}} = \frac{q \cdot d}{\|q\| \|d\|} \]

Edit Distance

Clustering
Course Middleware

Message-oriented Middleware / Asynchronous

Data Integration

Transaction Processing Application Architecture

<bib>
  <book title="a"
    <author name="x"/>
    <author name="y"/>
  </book>
  <book title="b">
    <author name="x"/>
    <author name="y"/>
  </book>
</bib>
Course Recent Developments for Data Models

Data Analysis via (Multidimensional) Data Cubes

User-defined Functions and Stored Procedures

CREATE FUNCTION distance
(loc1 VARCHAR(50),
loc2 VARCHAR(50))
RETURNS INTEGER ...;

Window Queries

Temporal Data Models

<table>
<thead>
<tr>
<th>ENo</th>
<th>EDept</th>
<th>Start</th>
<th>End</th>
</tr>
</thead>
<tbody>
<tr>
<td>22217</td>
<td>3</td>
<td>d01</td>
<td>d04</td>
</tr>
<tr>
<td>22217</td>
<td>4</td>
<td>d05</td>
<td>d07</td>
</tr>
<tr>
<td>22217</td>
<td>3</td>
<td>d08</td>
<td>d012</td>
</tr>
</tbody>
</table>
Projects, Seminars, Theses

• **DB Project (every summer)**
  - Creation of a DB, development of an application, ORDBS (user-defined functions, stored procedures), **application development for data warehousing/OLAP**

• **IS Project (every winter)**
  - Implementation of a **web search engine**: multi-threaded web crawler, indexing of HTML pages, search (CLI and HTML UI), ranking models, link analysis (PageRank), image search, ad placement, duplicate detection, ....

• **Seminar**
  - Offered every semester
  - Topics depend on novel, interesting publications in our field or are centered around a specific topic, like modern hardware or Big Data.

• **Theses Offers**
  - Wide variety along the contents of the specialized courses. Pick your interests during the courses and get in contact with us!
<table>
<thead>
<tr>
<th>Projects</th>
<th>Lectures</th>
<th>Seminar, 4CP, (every semester)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DB Schema Design and Programming</td>
<td>Middleware for Heterogen. and Distributed IS</td>
<td>Building a Web Search Engine</td>
</tr>
<tr>
<td>8 CP (every summer)</td>
<td>8CP (every winter)</td>
<td>8 CP (every winter)</td>
</tr>
<tr>
<td>Building a Web Search Engine</td>
<td>Distributed Data Management</td>
<td>(every other summer, next '21)</td>
</tr>
<tr>
<td>8 CP (every winter)</td>
<td>Information Retrieval and Data Mining</td>
<td>4CP (every other summer, next '21)</td>
</tr>
<tr>
<td>(every summer)</td>
<td>4CP (every other summer, next '20)</td>
<td>(every other summer, next '21)</td>
</tr>
<tr>
<td>Database Systems</td>
<td>Database Systems</td>
<td>(every winter)</td>
</tr>
<tr>
<td>8CP (every winter)</td>
<td>8CP (every winter)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(every other summer)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(every semester)</td>
<td></td>
</tr>
</tbody>
</table>

mandatory
Databases and Information Systems Lab
Prof. Michel
https://dbis.cs.uni-kl.de/

Heterogeneous Information Systems Lab
Prof. Deßloch
http://www.lgis.cs.uni-kl.de/

Next Courses

Winter 19/20 Database Systems, Middleware, Seminar
Summer 20 Recent Developments, Schema Design Project, Seminar
Winter 20/21 Database Systems, Middleware, Search Engine Project, Seminar
Summer 21 Distributed Data Management, Information Retrieval and Data Mining, Schema Design Project, Seminar