Digi-Dak Database Project (DDDP)
Dienstags, 15-17 Uhr, Raum 29/335

Prof. Gunter Saake, Sandro Schulze, and Martin Schäler

October 19, 2010
Overview

- Introduction to AG DB (Lectures, Research)
- Short introduction to Digi-Dak
- Concepts of DDDP
- Overview of project topics & forming project teams
- Course of action (milestones, presentations)
- How to perform literature research?
Overview

- Introduction to AG DB (Lectures, Research)
- Short introduction to Digi-Dak
- Concepts of DDDP
- Overview of project topics & forming project teams
- Course of action (milestones, presentations)
- How to perform literature research?
- Further lectures:
  - Academic writing (2 lectures)
Organization
AG Databases: Research Fields

- Database Technologies
  - Integration of Information Systems
  - Tuning/Self-Tuning of DBMS
  - Multimedia Databases
  - Tailor-made Data Management
- Software Engineering
  - Feature-Oriented Software Development (FOSD)
  - Adaptive Information Systems
- Complex and Embedded Information Systems
  - Embedded and Automotive Systems
  - Interoperability
- Research projects:
  http://wwwiti.cs.uni-magdeburg.de/iti\-db/forschung/index.php#projekte
AG Databases: Lectures (I)

Winter Term

Bachelor
- Datenbanken I
- EPMD
- Digi-Dak Database Project
- Seminar: Empirical Software Engineering
- Seminar: Datenqualität

Master
- Transaktionsverwaltung
- Distributed Data Management (DDM)
- Data Warehouse Technologies (DWT)
- Erweiterte Programmierkonzepte für maßgeschneiderte Datenhaltung (EPMD)
- Digi-Dak Database Project
- Filmseminar
AG DB: Lectures (II)

Summer Term

Bachelor

- Spezifikationstechnik
- Databases 2
  (implementation techniques)
- Seminar: DM in eingebetteten Systemen

Master

- Advanced Topics in Databases
- Advanced Database Models
- Student Conference (SCSEDB)

General

- Different Software Projects:
  http://wwwiti.cs.uni-magdeburg.de/iti\_db/study/index.php#praktika
- Bachelor and Master Thesis
Digi-Dak: Introduction

- Research project funded by Federal Ministry for Education and Research (BMBF)
- Dig-Dak ⇒ Digital Dactyloscopy
- Research on pattern matching techniques of (digital) fingerprints
- Different fields of research: signal/image processing, IT security/forensics, database systems
- Different scenarios (for preventative/forensic processes):
  - Aging detection
  - Fingerprint overlapping
- More information:
  http://omen.cs.uni-magdeburg.de/digi-dak
Bachelor

► **Einordnung:** WPF FIN SMK (Schlüssel- und Methodenkompetenzen)
► 5 CP = 150h ⇒ 56h presence time (4 SWS) + 94h autonomous work

**Master**

► **Einordnung:** Wissenschaftliches Teamprojekt
► 6 CP = 180h ⇒ 56h presence time (4 SWS) + 124h autonomous work

*Grade at the end of the course for the whole project team*
Concepts & Content
Project Topics

1. Secure Database Infrastructure
2. Data Provenance
3. Database Tuning regarding high-dimensional data

More information later in this lecture ...
Lecture, Meetings & Presentation

Lecture & Presentation

▶ Time/Place: tuesdays, 3 to 5 p.m., G29 - room 335
▶ Lectures with content of course → all
▶ Presentation of main milestones (see gantt chart) → each project team

Meetings (Exercise)

▶ Individual for each project team
▶ Time and room to be agreed in project teams!
▶ Presentation of all intermediate results/milestones (informal)
▶ Discussion, discussion, discussion . . .
The Idea . . .

Role-playing game . . . Imagine
You

We
Role-playing game... Imagine

You

▶ ... are an upcoming project team
▶ ... are searching for innovative DB solutions
▶ Research is your key to success

We
The Idea . . .

*Role-playing game. . . Imagine*

*You*

- . . . are an upcoming project team
- . . . are searching for innovative DB solutions
- Research is your key to success

*We*

- . . . are the customer
- . . . want the best solution that gives us an unique selling proposition
- . . . want you to provide evidence of your scientific capabilities
Acquired skills, *specific to research*

- Performing literature research
- Understanding and structured review of scientific work
- Autonomous, solution-based reasoning on research task (e.g., finding alternative solutions)
- How to ask? How to adapt a task (extend/reduce)?
- Academic writing
Objectives & Qualification (II)

**Acquired skills, always needed**

- Team management
- Project and time scheduling
- Presentation of results
- Flexibility regarding changing conditions
- Reasoning about solutions ("Why is this the best/not adequate..."")
Progress of Course

Deliveries

- 6 milestone presentations (*main milestones*)
- Each team member has to present at least once
- Reporting of (sub) milestones in exercises/meetings
- Written paper about literature research (technical report)
- Management report
- Prototypical implementation
Grading Criteria (I)

Technical Report

- Delivery of report in sufficient time
- Number of pages
- Quality/Quantity of Structure and Evaluation
- Quality/Quantity of literature research
- Own contribution
Grading Criterias (II)

Management Report

- Description of project realization (timeline, milestones)
- Separation of roles and contributions of single team members
- Meeting protocols
- Self-evaluation of member and group work (strengths, weaknesses)
Grading Criteria (III)

Presentation & Discussion

- Quality of scientific presentation (structure, references, time)
- Assessment regarding the content (e.g., results of particular milestones)
- Participation of discussion

Organization

- Strictness
- Communication (just-in-time answers, satisfying time constraints)
- Self-organization (Sharing tasks, internal reporting of current state-of-work, dealing with problems)
- Autonomous working
Teams with 3 to 6 students
Every task can be chosen once
Projects
  - Theoretical part
    - State of the art
    - New ideas
  - Practical part
    - DBMS PostgreSQL 8.4
    - Prototypical implementation
Prevent unwanted Access to the DB

- Focus: theoretical part
- Theoretical part
  - State of the art
  - New ideas
- Practical part
  - DBMS PostgreSQL 8.4
  - Show basic idea
  - Security test framework
**Theoretical part**

- Which security problems do exist? (Classification: Attacks, Leaks, ...)
- Threat model
  - Where?
  - Who?
- Making the DB secure
  - Generally
  - Especially in PostgreSQL 8.4
- Evaluation of Postgre’s security mechanism
- Literature
  - Handbook of DB Security
  - PostgreSQL Documentation
Topic 1 - Secure Data Base Infrastructure

Practical part

▶ Demonstrator
  ▶ PostgreSQL 8.4
  ▶ Client server communication

▶ Security test framework
  ▶ Collection of Attacks

▶ Intra team competition
  ▶ E.g. sub-teams with two students
  ▶ Team A(ttacker) tries to get access to the DB controlled by team B
  ▶ Switch the roles after each successful attack
  ▶ be creative ...
Topic 2 - Reliable Data Provenance

What is/are the sources of my data and what transactions have been performed?

► Focus practical part: Own ideas
► Theoretical part
  ► State of the art
  ► New ideas
► Practical part
  ► DBMS PostgreSQL 8.4
  ► Implementing existing and (own) new ideas
  ► Circumvention of provenance mechanism
**Topic 2 - Reliable Data Provenance**

*Theoretical part*

- Which approaches do exist? (classification)
  - Type of Data (structured, raw)
  - Classification criteria (integrity, authenticity, ...)
  - Mechanisms

- Literature
  - Paper Collection
  - Handbook of DB Security
**Practical part**

- Demonstration
  - PostgreSQL 8.4
  - Several clients and transformations
  - Invent 3 scenario for structured, raw and mixed data
  - Identify origin(s) and transformation process

- Iterative implementation
  - Intra team competition
  - Illustrate and implement existing/own approaches
  - (1) Circumvent them!
  - (2) Improve the Approach
  - (3) Goto to (1)

- Document *every* result
How to find all "similar" Data Sets as fast as possible?

- Focus: practical part - Experiments
- Theoretical part
  - Short general information
  - What is different in high dimensional data?
- Iterative implementation
  - Demonstrator: Experiments
  - DBMS PostgreSQL 8.4
**Scenario**

- DB with large number \((> k \times 65.536)\) of images
- Image: \(i \times i\) \((i=2^n)\) 8bit gray values
- Input: Image
- **Task (a) - Exact matches**
  - Determine whether the image is in the DB
  - Minimize the HDD accesses: goal one access
  - Output: ID(s) of the image(s) or No
- **Task (b) - Nearest neighbors**
  - Calculate the \(n\) nearest neighbors by some well defined distance measure
  - Upper bound of HDD accesses (e.g. read not more than 20% of the whole data)
  - Output: List of candidates
- **Document every result**
Topic 3 - DB Tuning in high dimensional Data Sets

*Theoretical part*

- Short general infos
- Tuning in high dimensional data
  - What is different?
  - Consider different query types as read only, exact match, nearest neighbors etc.
  - Problem: Dimensional curse
- Tuning mechanism in PostgreSQL 8.4
  - Which mechanism do exist?
  - Classification: logical / physical
  - Which promising approaches cannot be used in PostgreSQL?
Practical part

- Demonstrator in for both tasks
- Task (a) - Exact match
  - Chose different index structures as B-trees derivatives, hashes, multi dimensional indexes ... (see DB II)
  - Reference: Full table scan
- Task (b) - Nearest neighbors
  - Define different distance measurements and implement them in the DB
  - Reference: Difference per bit with full table scan
- Creating the measurement environment
  - Reproducible, comprehensible results
  - Measurement or cost model (query execution plans)
Practical part cont.

- Iterative approach: Start (both tasks)
  - Start with 2×2 Pixel
  - Reference: Full table scan
  - Compare speed up by indexes
- Create a test plan
- Increase amount of data (4×4, 8×8, etc.)
- Increase the number of images(k)
- Document every result
Task & Time Management

Task Management

▶ *Main milestones* have to be finished in time
▶ *(Sub) milestones* are less strict (but don’t be sloppy)
▶ Pre-defined work packages ⇒ each project team
  ▶ . . . defines sub work packages
  ▶ . . . determines responsibilities for these packages
    (divide & conquer)

Time Management

▶ Planning of periods
▶ Regarding capacities and resources
▶ Considering other tasks and activities
▶ Reporting of delays immediately to project members!
Role Management

- Possible roles: team leader, design, implementation, testing, writing, ...
- Delegate for important roles/work packages
- Assignment of (sub) tasks to role for each milestone
Individual Criteria Examples

- Is open and forthcoming in discussions
- Is democratic (i.e. willingly accepts team decisions)
- Is punctual for team meetings
- Is reliable in finishing assigned work
- Is creative and full of ideas in respect to (a) team organization and b) technical approaches
- Communicates clearly
- Contributes significantly to the project (not only in spending time)
- Takes the initiative when appropriate
- Is cooperative
- Is prepared to spend the time necessary to complete his/her tasks
- Is prepared to take responsibility
Literature Research
How to Perform Literature Research

Efficient literature research requires

- Knowledge of *Where* to search
- Knowledge of *How* to search
- Finding adequate search terms
- Structured review of papers
- Knowledge of how to find information in papers
Where to Search (I)

- Different websites available that provide large literature databases

1. Google Scholar: http://scholar.google.de/
   - Key word and concrete paper search
   - Often, PDFs are provided

2. DBLP: http://www.informatik.uni-trier.de/~ley/db/
   - Search for keyword, conferences, journals, author(s)
   - BibTex and references to other websites

3. Citeseer: http://citeseerx.ist.psu.edu/about/site
   - keyword, fulltext, author, and title search
   - BibTex and (partially) PDFs are provided
Where to Search (II)

- Publisher sites are also a suitable target
- ACM Digital Library: http://portal.acm.org/dl.cfm
  - Keyword, author, conference/literature (proceedings), and title search
  - Bibtex, mostly PDFs and other information are provided
  - Similar to ACM, but only few PDFs
  - Extended access within university network
- Springer: http://www.springerlink.de/
  - Similar to previous
  - Extended access within university Network
- Further search possibilities: on author, research group or university sites
How to Search

Some hints to not get lost in the jungle

▶ Use distinct keywords (fingerprint vs. fingerprint data)
▶ Keep keywords simple (at most three words)
▶ Otherwise, search for whole title
▶ Read abstract (and maybe introduction) ⇒ decision for relevance

First insights

▶ Read abstract, introduction and background/related work (coarse-grained) to
  ▶ ... get a first idea of the approach
  ▶ ... find other relevant papers
Information Retrieval

Finding the required information

- Read the paper carefully
- Omit formal parts/sections
- Try to classify (core idea, main characteristics) ⇒ develop classification/evaluation in mind
- Understand the big picture
- Make notes
- Do NOT translate each sentence