Scientific Project: Databases for Multi-dimensional Data, Genomics, and modern Hardware

David Broneske, Sebastian Dorok, Andreas Meister, Veit Köppen, Martin Schäler, Gunter Saake

April 13, 2015
Overview

- Introduction to AG DBSE (Lectures, Research)
- Concepts
- Overview of project topics & forming project teams
- Course of action (milestones, presentations)
- How to perform literature research?
Overview

- Introduction to AG DBSE (Lectures, Research)
- Concepts
- Overview of project topics & forming project teams
- Course of action (milestones, presentations)
- How to perform literature research?
- Further lectures:
  - Academic writing (2-3 lectures)
Organization
AG Databases & Software Engineering: Research Fields

Database Technologies
- Integration of Information Systems
- Cloud Systems
- Multimedia Databases
- Data Management on Modern Hardware
- Databases for Genomics
- Tailor-made Data Management

Software Engineering
- Feature-Oriented Software Development (FOSD)
- Testing & Verification of SPLs
- Multi Software Product Lines
- Refactoring

Research projects:
http://wwwiti.cs.uni-magdeburg.de/iti_db/forschung/index.php#projekte
Winter Term
Bachelor
► Datenbanken I
► Implementierungs-techniken für Software-Produktlinien

Master
► Transaktionsverwaltung
► Distributed Data Management
► Data Warehouse Technologies
► Implementierungs-techniken für Software-Produktlinien
► Filmseminar
AG DB: Lectures (II)

Summer Term

Bachelor

- Database Concepts
- Databases 2
  (implementation techniques)

Master

- Advanced Topics in Databases
- Advanced Database Models
- Student Conference

General

- Scientific Project
- Software Projects
- Bachelor and Master Thesis: http://wwwiti.cs.uni-magdeburg.de/iti_db/study/index.php#theses
Scientific Project: Course Grading

Bachelor

- **Module:** WPF FIN SMK (Schlüssel- und Methodenkompetenzen)
- 5 CP = 150h ⇒ 42h presence time (3 SWS) + 108h autonomous work

Master

- **Module:** Scientific Team Project
- 6 CP = 180h ⇒ 42h presence time (3 SWS) + 138h autonomous work

*Grade at the end of the course for the whole project team*
Scientific Project: Course Grading II

- Weighting the Grade:
  - Presentations: 30%,
  - Implementation: 30%,
  - Paper: 30%,
  - Soft Skills: 10%

- Binding registration: Second Milestone
<table>
<thead>
<tr>
<th></th>
<th>Monday</th>
<th>Tuesday</th>
<th>Wednesday</th>
<th>Thursday</th>
<th>Friday</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong></td>
<td>06.04.15</td>
<td>07.04.15</td>
<td>08.04.15</td>
<td>09.04.15</td>
<td>10.04.15</td>
</tr>
<tr>
<td><strong>Aca-I</strong></td>
<td>13.04.15</td>
<td>14.04.15</td>
<td>15.04.15</td>
<td>16.04.15</td>
<td>17.04.15</td>
</tr>
<tr>
<td></td>
<td>20.04.15</td>
<td>21.04.15</td>
<td>22.04.15</td>
<td>23.04.15</td>
<td>24.04.15</td>
</tr>
<tr>
<td><strong>Aca-II</strong></td>
<td>27.04.15</td>
<td>28.04.15</td>
<td>29.04.15</td>
<td>30.04.15</td>
<td>01.05.15</td>
</tr>
<tr>
<td><strong>MS-I</strong></td>
<td>04.05.15</td>
<td>05.05.15</td>
<td>06.05.15</td>
<td>07.05.15</td>
<td>08.05.15</td>
</tr>
<tr>
<td></td>
<td>11.05.15</td>
<td>12.05.15</td>
<td>13.05.15</td>
<td><strong>14.05.15</strong></td>
<td><strong>15.05.15</strong></td>
</tr>
<tr>
<td><strong>MS II</strong></td>
<td>18.05.15</td>
<td>19.05.15</td>
<td>20.05.15</td>
<td>21.05.15</td>
<td>22.05.15</td>
</tr>
<tr>
<td></td>
<td><strong>25.05.15</strong></td>
<td>26.05.15</td>
<td>27.05.15</td>
<td>28.05.15</td>
<td>29.05.15</td>
</tr>
<tr>
<td><strong>MS III</strong></td>
<td><strong>01.06.15</strong></td>
<td>02.06.15</td>
<td>03.06.15</td>
<td>04.06.15</td>
<td>05.06.15</td>
</tr>
<tr>
<td></td>
<td><strong>08.06.15</strong></td>
<td><strong>09.06.15</strong></td>
<td><strong>10.06.15</strong></td>
<td><strong>11.06.15</strong></td>
<td><strong>12.06.15</strong></td>
</tr>
<tr>
<td></td>
<td>15.06.15</td>
<td>16.06.15</td>
<td>17.06.15</td>
<td>18.06.15</td>
<td>19.06.15</td>
</tr>
<tr>
<td></td>
<td>22.06.15</td>
<td>23.06.15</td>
<td>24.06.15</td>
<td>25.06.15</td>
<td>26.06.15</td>
</tr>
<tr>
<td><strong>MS Final</strong></td>
<td><strong>06.07.15</strong></td>
<td><strong>07.07.15</strong></td>
<td><strong>08.07.15</strong></td>
<td><strong>09.07.15</strong></td>
<td><strong>10.07.15</strong></td>
</tr>
</tbody>
</table>
Scientific Project: Milestones

- Milestone I - Topic, schedule, and team presentation & first results of literature research
- Milestone II - Concept & additional literature research
- Milestone III - Implementation & evaluation setup
- Milestone IV - Final presentation (wrap-up + evaluation results)
Concepts & Content
Lecture, Meetings & Presentation

*Lecture & Presentation*

- Time/Place: Monday, 11:15-12:45, G29 - room K059
- Lectures with content of course → all
- Presentation of *main milestones* (see time table)
  → 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 . . .
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
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 a unique selling proposition
- ... want you to provide evidence of your scientific capabilities
Acquired skills, specific to research

- Performing literature research
- Understanding and structured reviewing 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
- 4 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
Deliveries and Grading (I)

**Technical Report**

- Delivery of report at a given time (deadline)
- Quality/Quantity of literature research
- Number of pages
- Quality of paper structure and evaluation
- Own contribution
Deliveries and Grading (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)
Deliveries and Grading (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
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, developer, researcher, . . .
- work together vs. responsibilities: design, implementation, testing, writing, . . .
- Delegate for important roles/work packages
- Assignment of (sub) tasks to role for each milestone
Topic & Project Teams

- Teams with 3 to 5 students (depends on the task)
- Every task can be chosen once
- **Projects**
  - Theoretical part
    - State of the art
    - New ideas
  - Practical part
    - Usually JAVA or C++
    - Prototypical implementation
    - Evaluation part
Research project, started in April 2012
CoGaDB ⇒ Column-oriented GPU-accelerated DBMS
Research on co-processor-accelerated data management
Different fields of research: efficient algorithms for (co-)processors, in-memory databases, query optimization
Especially designed for OLAP queries
More information:
http://cogadb.cs.tu-dortmund.de/wordpress/
Topic 1.1 - Parallel Grouping in CoGaDB

Intro

- Groupings operate on a huge amount of data
  → Highly data intensive operation
- For 3-5 Bachelor/Master Students

Task

- What are common techniques? (sort-based vs. hash-based)
- Prototypical implementation of at least 2 techniques/variants in CoGaDB using C++
- Tune algorithms to performance (e.g., using a profiler)
- Evaluate their performance and draw conclusions
Topic 1.2 - Optimized Scans in CoGaDB

Intro
- Scans are essential in today's database systems!
- Many proposed code optimizations
- Current scans in CoGaDB are tuned using:
  - Branch Free Code
  - Loop Unrolling
  - SIMD Acceleration
  - Parallelization
- For 3-5 Bachelor/Master Students

Task
- Literature Research: Collect further code optimizations (e.g., bloom filters)
- Prototypical implementation of further optimized scans in CoGaDB using C++
- Evaluate the performance for them w.r.t. existing variants
**Intro**

- Scans are essential in today's database systems!
- Many proposed code optimizations
- Our current Scans use SSE 4.2 intrinsics
  - What about AVX?
- For 3-5 Bachelor/Master Students

**Task**

- Literature Research: Collect literature about SIMD Scans
- Prototypical implementation of scans using AVX-instructions in CoGaDB
- Evaluate the performance for them w.r.t. existing variants
Topic 1.4 - Join Order Optimization in CoGaDB

Intro

- Join Order Optimization needed for efficient query processing
  \(\rightarrow\) NP-hard problem
- For 3-5 Bachelor/Master Students

Task

- What are common techniques? (Dynamic Programming, Genetic algorithms, ...)
- Prototypical implementation of at least 2 techniques/variants in CoGaDB using C++
- Tune algorithms to performance (e.g., using a profiler)
- Evaluate their performance and draw conclusions
Amount of available genome data increases rapidly
Analysis of big data is required
... while data management challenges become more important than ever
State-of-the-art: Command-line driven and flat-file based :(  
Topics are for 3-5 students (rather master)
Intro

- Instead of reinventing the wheel, just use an existing DBMS, but ... Which one to choose?
  - RDBMSs provide well-established and proven data management capabilities
  - NoSQL Systems seem to scale better for Big Data Analytics
Topic 2.1 - DBMSs for Genome Data Management

Intro

▶ Instead of reinventing the wheel, just use an existing DBMS, but ... Which one to choose?
  ▶ RDBMSs provide well-established and proven data management capabilities
  ▶ NoSQL Systems seem to scale better for Big Data Analytics

Task

▶ Literature Research: How are DBMSs used for managing genome data? What are typical use cases?
▶ Implement a use case using an RDBMS (e.g., MySQL) and a NoSQL DBMS (e.g., Cassandra)
▶ Evaluate your approaches and draw conclusions
Topic 2.1 - DBMSs for Genome Data Management

Intro
▶ Instead of reinventing the wheel, just use an existing DBMS, but ... Which one to choose?
  ▶ RDBMSs provide well-established and proven data management capabilities
  ▶ NoSQL Systems seem to scale better for Big Data Analytics

Task
▶ Literature Research: How are DBMSs used for managing genome data? What are typical use cases?
▶ Implement a use case using an RDBMS (e.g., MySQL) and a NoSQL DBMS (e.g., Cassandra)
▶ Evaluate your approaches and draw conclusions

We have
▶ NOSQL implementations and
▶ CoGaDB implementation to reuse or extend
Topic 2.2 - Variant Calling inside Database Systems

Intro

- Variations in genomes matter → body characteristics but also susceptibility for diseases
- Different methods available to determine (call) variants
Topic 2.2 - Variant Calling inside Database Systems

Intro

▶ Variations in genomes matter → body characteristics but also susceptibility for diseases
▶ Different methods available to determine (call) variants

Task

▶ Literature Research: Variant calling approaches
▶ Prototypical implementation of a two different variant calling approaches in CoGaDB
▶ Evaluate your implementation w.r.t. efficiency and performance
▶ Compare against state-of-the-art tools
Intro

- Variations in genomes matter → body characteristics but also susceptibility for diseases
- Different methods available to determine (call) variants

Task

- Literature Research: Variant calling approaches
- Prototypical implementation of a two different variant calling approaches in CoGaDB
- Evaluate your implementation w.r.t. efficiency and performance
- Compare against state-of-the-art tools

We have

- A relational database schema for aligned genome data
- Data loading tools for genome data in CoGaDB
Finding your Team

Topics:

- Topic 1.1 - Parallel Grouping in CoGaDB
- Topic 1.2 - Optimized Scans in CoGaDB
- Topic 1.3 - SIMD Scans in CoGaDB
- Topic 1.4 - Join Order Optimization in CoGaDB
- Topic 2.1 - DBMSs for Genome Data Management
- Topic 2.2 - Variant Calling inside Database Systems
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
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
Finding your Team

Topics:

- Topic 1.1 - Parallel Grouping in CoGaDB
- Topic 1.2 - Optimized Scans in CoGaDB
- Topic 1.3 - SIMD Scans in CoGaDB
- Topic 1.4 - Join Order Optimization in CoGaDB
- Topic 2.1 - DBMSs for Genome Data Management
- Topic 2.2 - Variant Calling inside Database Systems