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

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

October 16, 2014
Overview

▶ Introduction to AG DBSE (Lectures, Research)
▶ Short introduction to Digi-Dak+
▶ 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)
- Short introduction to Digi-Dak+
- 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
AG Databases: Lectures (I)

Winter Term

Bachelor
- Datenbanken I
- Implementierungstechniken für Software-Produktlinien
- Seminars

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

Summer Term

Bachelor

- Databases 2
  (implementation techniques)
- Seminars

General

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

Master

- Advanced Topics in Databases
- Advanced Database Models
- Student Conference
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>13.10.14</td>
<td>14.10.14</td>
<td>15.10.14</td>
<td>16.10.14</td>
<td>17.10.14</td>
</tr>
<tr>
<td><strong>Aca-II</strong></td>
<td>27.10.14</td>
<td>28.10.14</td>
<td>29.10.14</td>
<td>30.10.14</td>
<td><strong>31.10.14</strong></td>
</tr>
<tr>
<td><strong>MS-I</strong></td>
<td>03.11.14</td>
<td>04.11.14</td>
<td>05.11.14</td>
<td>06.11.14</td>
<td>07.11.14</td>
</tr>
<tr>
<td></td>
<td>10.11.14</td>
<td>11.11.14</td>
<td>12.11.14</td>
<td>13.11.14</td>
<td>14.11.14</td>
</tr>
<tr>
<td><strong>MS II</strong></td>
<td>17.11.14</td>
<td>18.11.14</td>
<td>19.11.14</td>
<td>20.11.14</td>
<td>21.11.14</td>
</tr>
<tr>
<td></td>
<td>24.11.14</td>
<td>25.11.14</td>
<td>26.11.14</td>
<td>27.11.14</td>
<td>28.11.14</td>
</tr>
<tr>
<td><strong>Winter Break</strong></td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td><strong>MS IV</strong></td>
<td><strong>05.01.15</strong></td>
<td><strong>06.01.15</strong></td>
<td>07.01.15</td>
<td><strong>08.01.15</strong></td>
<td><strong>09.01.15</strong></td>
</tr>
<tr>
<td></td>
<td>12.01.15</td>
<td>13.01.15</td>
<td>14.01.15</td>
<td>15.01.15</td>
<td>16.01.15</td>
</tr>
<tr>
<td></td>
<td>19.01.15</td>
<td>20.01.15</td>
<td>21.01.15</td>
<td>22.01.15</td>
<td>23.01.15</td>
</tr>
<tr>
<td></td>
<td>26.01.15</td>
<td>27.01.15</td>
<td>28.01.15</td>
<td>29.01.15</td>
<td>30.01.15</td>
</tr>
<tr>
<td><strong>MS Final</strong></td>
<td><strong>02.02.15</strong></td>
<td><strong>03.02.15</strong></td>
<td><strong>04.02.15</strong></td>
<td><strong>05.02.15</strong></td>
<td><strong>06.02.15</strong></td>
</tr>
</tbody>
</table>
Scientific Project: Milestones

- Milestone I - Topic, schedule, and team presentation
- Milestone II - First results of literature research
- Milestone III - Concept & additional literature research
- Milestone IV - Implementation
- Milestone V - Final presentation
Concepts & Content
Lecture, Meetings & Presentation

*Lecture & Presentation*

- Time/Place: Thursday, 9:15-10:45, 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 . . .
The Idea . . .

*Role-playing game . . . Imagine*
*You*

*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
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
Objectives & Qualification (I)

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*

- 5 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
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 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, team member
- work together vs. responsibilities: design, implementation, testing, writing, ...
- Delegate for important roles/work packages
- Assignment of (sub) tasks to role for each milestone
Teams with 2 to 3 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
Digi-Dak(+): Introduction and Use Case

- Research project funded by Federal Ministry for Education and Research (BMBF)
- Digi-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 (e.g., age determination)
- Digi-Dak+ ⇒ Graduate program
- More information: http://omen.cs.uni-magdeburg.de/digi-dak
QuEval: Evaluation of high-dimensional index structures

- WiSe 2010/11 - First qualitative evaluation and overview
- SoSe 2011 - Quantitative evaluation
- WiSe 2011/12 - Distance metrics
- SoSe 2012 - Impact of data distribution
- WiSe 2012/13 - String similarity (EDBT contest)
Task characteristics

- 3 Students
- Rather Master
- Different Developments of *Approximating and Eliminating Search Algorithm*
  - What approaches are known?
  - State of the art and classification
  - What are the differences between the approaches, which domain is addressed?
  - Prototypical implementation/evaluation
- We have: Index Framework
Topic 1.1 - AESA, LAESA, ROAESA, AECA, TLAESA

Theoretical part

▶ Discussing state of the art
  ▶ Literature research → Collect approaches
  ▶ Is there any work at all w.r.t. overall evaluation of AESA?
  ▶ Identification of solutions for practical part
▶ What would be nice to have?
Practical part

- Prototypical implementation
  - Additional: INSERT and DELETE (concept)
- Evaluation (examples)
  - AESA
  - LAESA
  - ROAESA
  - TLAESA
- Analysis → Computation effect estimation.
Topic 1.2 - Analysis of R Extensions for DBMSs

Task characteristics

- 2-3 Students
- Rather Bachelor

Task

- $R$ = Project for Statistical Computing
- Is used for analysis
Topic 1.2 - Analysis of R Extensions for DBMSs

Analysis Goals

▶ State of the art
  ▶ What extensions for well-known DBMS exist?
  ▶ How to use functionality of R in a DBMS?
  ▶ Data mapping from DB to R
  ▶ Application scenarios
  ▶ Things you should know when using R extensions
Task characteristics

- 3 Students
- Rather Master
- Transformation from Model to SQL
  - What is ADAPT?
  - State of the art for DDL in Multidimensional Context
  - What are
  - Prototypical implementation/evaluation
- We have: ORACLE DB, we can provide access for MS Visio
**Theoretical part**

- Discussing of multidimensional Modeling
  - Catalog of ADPAT model elements
  - Identification of SQL commands
  - Mapping

- What would be nice to have?
Practical part

- Prototypical implementation
  - Additional: INSERT and DELETE (concept)
- Evaluation (examples)
  - TPC-H Model
  - Star-Schema Benchmark Model
  - Example from Literature
- Analysis → Improvements & Challenges
Topic 2 - CoGaDB

- 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://wwwiti.cs.uni-magdeburg.de/iti_db/research/gpu/cogadb/
Topic 2.1 - Parallel Grouping in CoGaDB

Intro

▶ Groupings operate on a huge amount of data
  → Highly data intensive operation
▶ For 2-3 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 2.2 - 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 2-3 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
Topic 2.3 - Generating Scan Variants

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 2-3 Bachelor/Master Students

Task

- Literature Research: Code generation methods (e.g., JIT Compilation)
- Implement concepts for scan variant generation for selected code optimizations
- Evaluate their generalizability for arbitrary operations
Topic 2.4 - Join Order Optimization in CoGaDB

Intro

▶ Join Order Optimization needed for efficient query processing
   → NP-hard problem
▶ For 2-3 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
Topic 3 - Genome Data Management

- 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 2-3 Bachelor/Master Students
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 3.1 - 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 3.1 - 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
▶ Access to an Oracle 12c instance
▶ CoGaDB to do more advanced stuff ;)

David Broneske et al. Scientific Project
Topic 3.2 - Referential Compression for Genome Data

Intro

- Amount of genome data to analyse increases rapidly
- Referential compression of genome data promises huge storage savings
Topic 3.2 - Referential Compression for Genome Data

Intro

▶ Amount of genome data to analyse increases rapidly
▶ Referential compression of genome data promises huge storage savings

Task

▶ Literature Research: Compression schemes with focus on genome data, especially referential compression
▶ Prototypical implementation of a referential compression scheme in CoGaDB
▶ Evaluate your implementation w.r.t. efficiency and perf.
▶ Compare against built-in compression schemes
Topic 3.2 - Referential Compression for Genome Data

Intro

- Amount of genome data to analyse increases rapidly
- Referential compression of genome data promises huge storage savings

Task

- Literature Research: Compression schemes with focus on genome data, especially referential compression
- Prototypical implementation of a referential compression scheme in CoGaDB
- Evaluate your implementation w.r.t. efficiency and perf.
- Compare against built-in compression schemes

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 - AESA, LAESA, ROAESA, AECA, TLAESA
- Topic 1.2 - Analysis of R Extensions for DBMSs
- Topic 1.3 - Multidimensional Modeling & DDL
- Topic 2.1 - Parallel Grouping in CoGaDB
- Topic 2.2 - Optimized Scans in CoGaDB
- Topic 2.3 - Generating Scan Variants
- Topic 2.4 - Join Order Optimization in CoGaDB
- Topic 3.1 - DBMSs for Genome Data Management
- Topic 3.2 - Referential Compression for Genome Data
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) $\Rightarrow$ develop classification/evaluation in mind
- Understand the big picture
- Make notes
- Do NOT translate each sentence
Finding your Team

Topics:

- Topic 1.1 - AESA, LAESA, ROAES, AECA, TLAESA
- Topic 1.2 - Analysis of R Extensions for DBMSs
- Topic 1.3 - Multidimensional Modeling & DDL
- Topic 2.1 - Parallel Grouping in CoGaDB
- Topic 2.2 - Optimized Scans in CoGaDB
- Topic 2.3 - Generating Scan Variants
- Topic 2.4 - Join Order Optimization in CoGaDB
- Topic 3.1 - DBMSs for Genome Data Management
- Topic 3.2 - Referential Compression for Genome Data