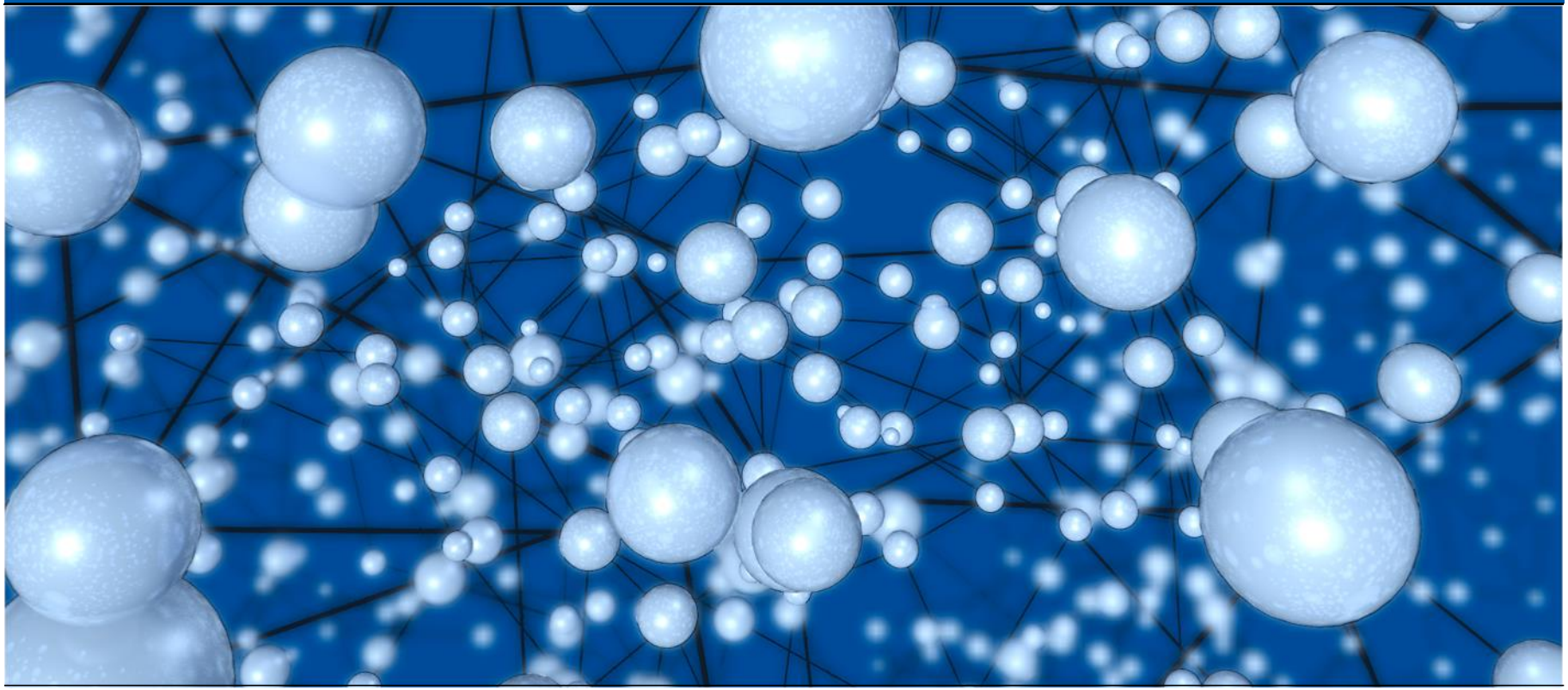
Database Engine Development Lab
SoSe 2015

TECHNISCHE UNIVERSITÄT DARMSTADT

Introduction & Organization

Alexander Frömmgen, Alejandro Buchmann



Project

TECHNISCHE UNIVERSITÄT DARMSTADT

- The goal of this lab is to develop and optimize certain features of a database engine.
- Depending on the first results and individual preferences, the advanced tasks can be chosen individually.

Basic knowledge about databases is required!

Sometimes, it makes sense to open a book!

What is a database engine?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Software component used by a database management system to create, read, update and delete data.

ORACLE DATABASE

- What are important design parts of a database engine?
  - Data structures (files, indexes, trees, hash tables, …)
  - Algorithms (scheduler, selection, projection, joins, …)
  - Storage hierarchy (disc, main memory, …)
  - …

MySQL

InnoDB,
MyISAM

Microsoft SQL Server

What you will do in this lab

TECHNISCHE UNIVERSITÄT DARMSTADT

- Build your own analytical database engine
  - Basic requirements for each team
  - Possible extensions depending on individual preferences

- What you will **not** do in this lab!

This course is **not** about writing applications that use database engines.

This course is **not** about modeling relational databases.

What you will do in this lab

TECHNISCHE UNIVERSITÄT DARMSTADT

- Build your own analytical database engine
  - Basic requirements for each team
    - Logical data structures
    - Physical data/storage structures
    - Buffer management
    - Query processing and operators (projection, selection, join, sort, aggregation...)
    - Scheduler
    - Cost functions and query optimizer
    - Performance measurement and optimization
  - Possible extensions depending on
    - Data compression
    - Inter- and intra-query parallelizati
    - Distribution of the query processing and the on different machines
    - Compilation of queries to native code/byte code at runtime
    - Index structures

We want to encourage you to implement optimizations inspired by publications, e.g. Just in-time Compilation for SQL Query Processing

What you will not do in this lab

TECHNISCHE UNIVERSITÄT DARMSTADT

- Dynamic insert operations
- Transaction and concurrency control
- User and access control

Programming language

TECHNISCHE UNIVERSITÄT DARMSTADT

- Whatever you like!
- E.g. C, C++, Java, Scala, Phyton, …

This might be a opportunity to learn and practise a new language!

Organization

TECHNISCHE UNIVERSITÄT DARMSTADT

- Teams of 3-4 students
- Every team develops its own database engine
  - analyses, designs, implements, tests, **tweaks***, and presents it
- Project divided into 6 or 7 phases
  - Each phase is 2 weeks in duration
- A meeting after every project phase

* This should be done systematically ;-)

What we expect

TECHNISCHE UNIVERSITÄT DARMSTADT

- A working and convincing database engine
- Report, source documentation, presentations
  - **eMail us report and slides the night before each meeting**
- Participation in programming, documentation, and presentations
  - Everyone in the group!
- Approximately 100-200h of work (1CP = 30h, 6CPs = 180h)
- **Ask for help if there are problems**

Meetings

TECHNISCHE UNIVERSITÄT DARMSTADT

- You deliver the lab report sketch
  - Deadline: midnight *before* meeting
- Each group presents
  - Current state of implementation (live demo!)
  - Approach taken
  - Progress
  - Problems
  - 4-5 slides (as PDF/PPT, delivered midnight before the meeting date)
- We present
  - The outline of the next phase
  - Relevant background theory where necessary
- We answer your questions

Mark Breakdown

TECHNISCHE UNIVERSITÄT DARMSTADT

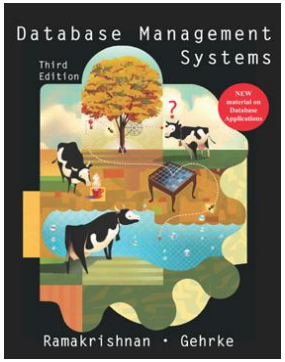
- **40% Implementation + Source Code Documentation**
  - Intermediate results count!
  - Coverage
  - Design
  - Correctness
  - Documentation
  - (Performance)
- **30% Meetings + Presentations**
  - Participation
  - Presentation (clarity and completeness)
  - Slides
- **30% Lab Report**
  - Overview of implementation and user manual
  - Documentation of development process
  - 10-15 pages total

Contact & Infrastructure

TECHNISCHE UNIVERSITÄT DARMSTADT

- Contact
  - froemmge@dvs.tu-darmstadt.de
- Course Homepage
  - http://www.dvs.tu-darmstadt.de/teaching/dbed/
- Our lab is available 24/7
  - Transponders are available from RBG service center C119

Literature



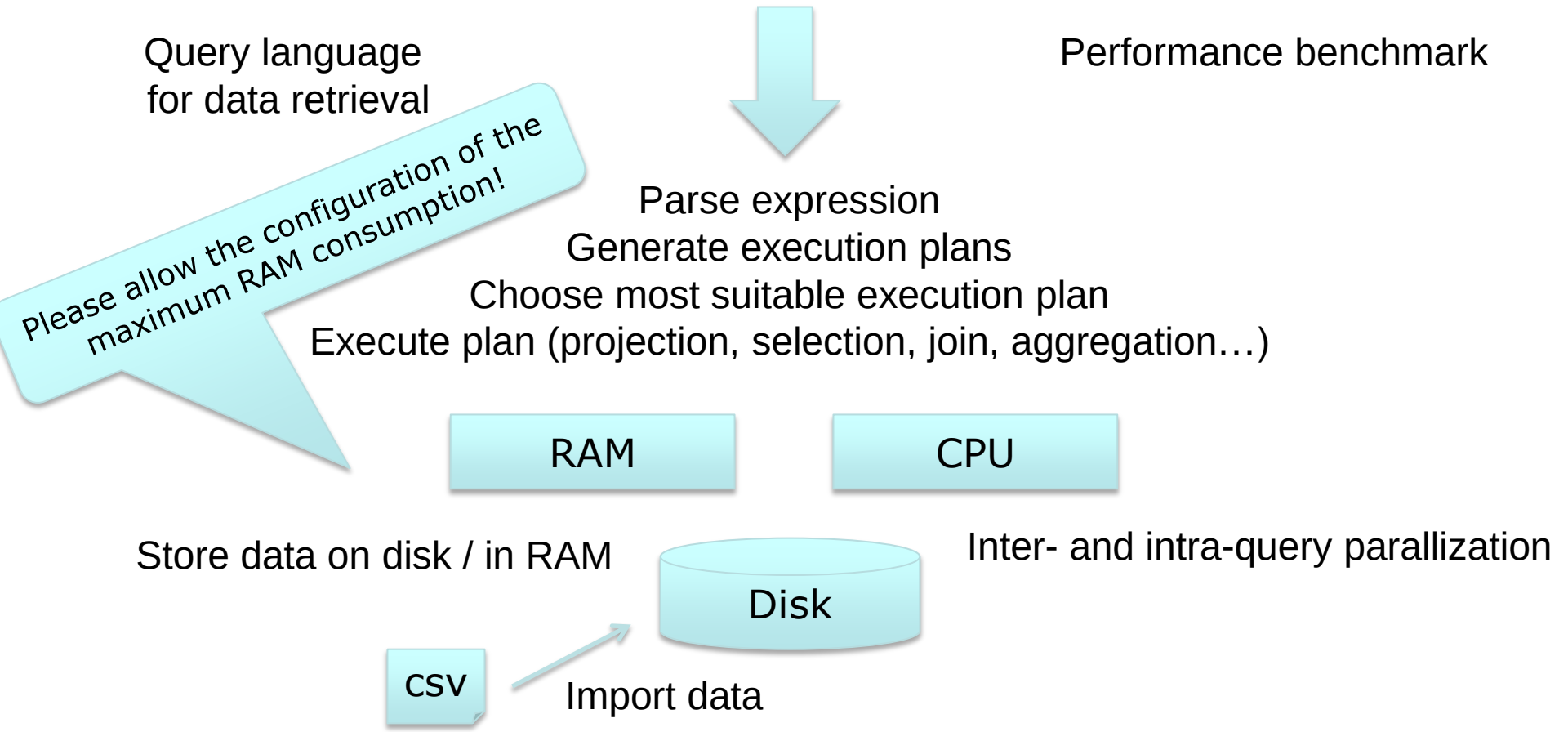
Database Management Systems
Ramakrishnan/Gehrke
3rd Edition

Database Systems II
http://www.dvs.tu-darmstadt.de/teaching/db2/

Business Intelligence and Data Warehousing
http://www.dvs.tu-darmstadt.de/teaching/bidw/

…

Blueprint of the Solution



Project Schedule

TECHNISCHE UNIVERSITÄT DARMSTADT

Summer Term (Vorlesungszeit)

April	**21.04.** (0)	Intro, query language, performance-benchmark
	28.04. (1)	Store data on disc / in RAM, execute selection
May	**12.05.** (2)	Execute projection, join, sort
	26.05. (3)	Performance benchmark
June	**09.06.** (4)	Optimizer / choose most suitable execution plan
	23.06. (5)	Advanced stuff
July	**14.07.** (6)	Advanced stuff
August	**04.08.** (7)	Final demonstration

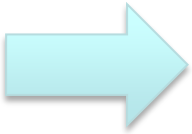
TECHNISCHE UNIVERSITÄT DARMSTADT

Phase 0
ANALYSIS AND DESGIN

Phase 1: Analysis & Design

TECHNISCHE UNIVERSITÄT DARMSTADT

- System architecture?
- Programming language?
- Design query language?
- Develop concrete test cases!
  - Correctness and performance testing
- How will you measure the performance of the system?
- First basic performance measurements!



Be prepared to present your results next week!

We provide you data

TECHNISCHE UNIVERSITÄT DARMSTADT

- As a csv file at the webpage (example1.zip)
- Will be extended in the next sessions!
- (StatisticId, Time, ExperimentId, NodeId) is unique

StatisticId	StatisticName	Time	ExperimentId	ExperimentName	NodeId	Unit	Value
109	managed/target density	2010-01-01 00:00:44	340	short-durable	37	bytes	704
109	managed/target density	2010-01-01 01:14:40	340	short-durable	19	seconds	32

Query language

TECHNISCHE UNIVERSITÄT DARMSTADT

- You do not have to use SQL!
- Some inspirations:
  - LINQ http://msdn.microsoft.com/de-de/library/bb397926.aspx
  - Some examples from the MSDN

```
List<Customer> customers = GetCustomerList();
var orderCounts = from c in customers
                  select new { c.CustomerID, OrderCount = c.Orders.Count() };

List<Product> products = GetProductList();
var categories = from p in products group p by p.Category into g
                  select new { Category = g.Key, AveragePrice = g.Average(p
=> p.UnitPrice) };

var categoryCounts = from p in products group p by p.Category into g
                  select new { Category = g.Key, ProductCount = g.Count() };
```

Query language

TECHNISCHE UNIVERSITÄT DARMSTADT

- You do not have to use SQL!
- Some inspirations:
  - Use Scala for a domain specific language

```
// open DB
val db = new SimDB("myDB");

// read data cube
val ds = new ExperimentDBSourceDataCube(db)

// make query
val ds2 = ds.filter().equals("Experiment", 320).avg("Node",
              "Time").rotate("Statistic").orderBy("Time")

// materialize
val result = ds2.evaluate();
```

Think about Performance
Measure Performance

TECHNISCHE UNIVERSITÄT DARMSTADT

- Concentrate on Performance!
- This implies a basic understanding of the relevant knobs!

- How long does it take to read/write a certain amount on SSD?
  - Sequential access
  - Random access

- How long does it take to read/write a certain amount of RAM?
  - Sequential access
  - Random access

➔ Are there buffers?

➔ Can you leverage multiple cores?

What's next?

TECHNISCHE UNIVERSITÄT DARMSTADT

- Meeting in **A213**, 29.04.14, 13:30
- You will
  - present your results
- We will give you
  - feedback on your results & implementation idea

The Sorting Hat

TECHNISCHE UNIVERSITÄT DARMSTADT

- Choose your Team.

DVS