HPI Hasso Plattner Institut

Digital Engineering · Universität Potsdam

Trends and Concepts in Software Industry I

Goals

Deep technical understanding of trends and concepts in enterprise computing, esp. main-memory-centric data management on modern hardware, cloud-native development and their impact on applications.

- Foundations of database storage techniques and operators
- Characteristics of enterprise applications and systems
- Trends in enterprise computing (e.g., cloud platforms)
- Hands-on exercises and experiments

Block Week

- General information
  - When: September (presumably) (or July)
  - Lectures given by Prof. Plattner
  - Additional lectures by guests from industry
  - Discussions about open questions in enterprise computing are a vital part of the lecture!

- Focus areas
  - Principles of in-memory databases
  - Characteristics of modern enterprise systems
  - Influence of cloud-native development
  - Trends in enterprise computing

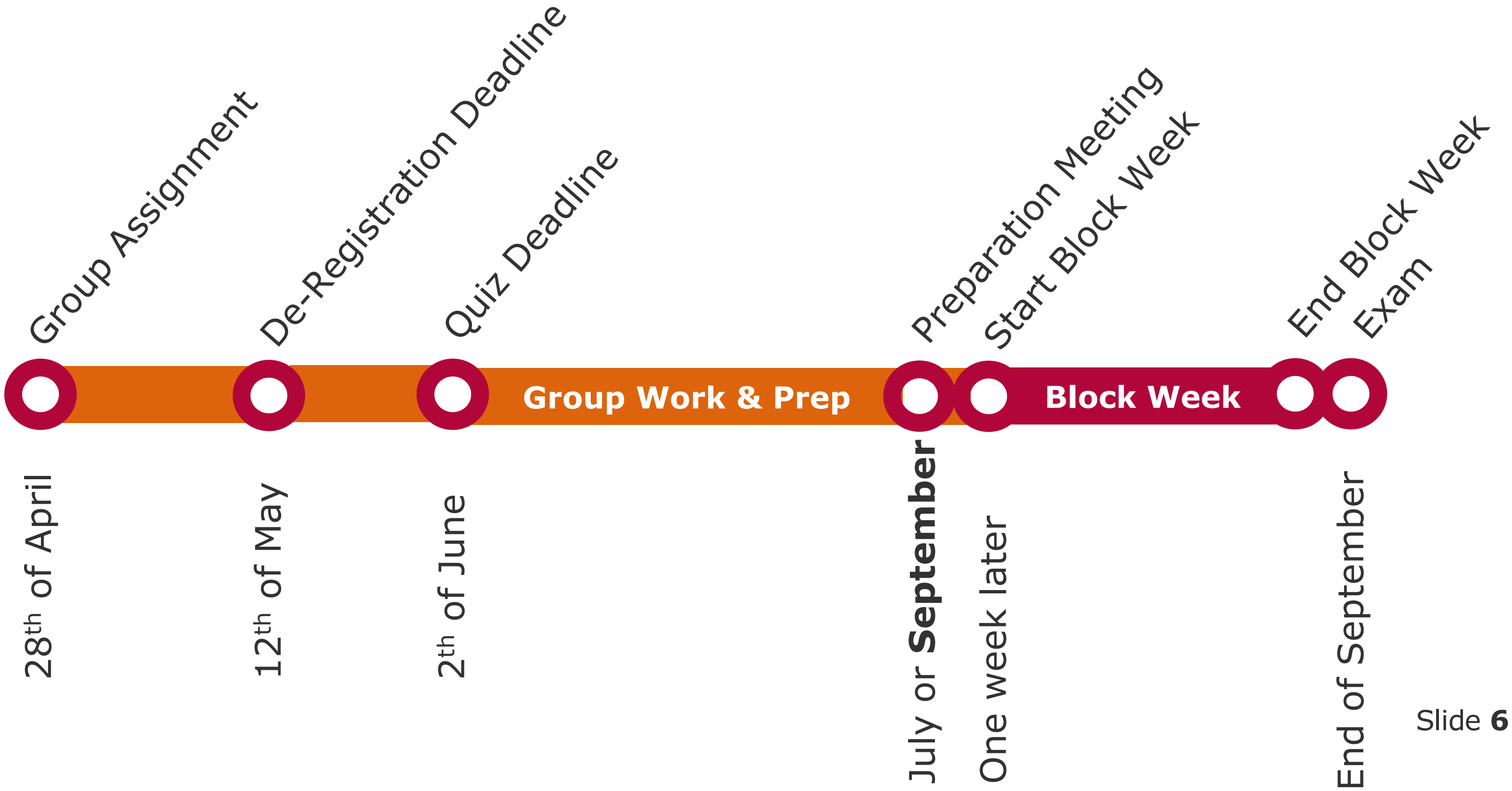
General Information

- 6 ECTS points
- Latest enrollment: 22nd of April 2020
- Modules
  - IT-Systems Engineering MA
    - ITSE-Analyse
    - ITSE-Entwurf
    - ITSE-Konstruktion
    - ITSE-Maintenance
    - BPET-Konzepte und Methoden
    - BPET-Spezialisierung
    - BPET-Techniken und Werkzeuge
    - SAMT-Konzepte und Methoden
    - SAMT-Spezialisierung
    - SAMT-Techniken und Werkzeuge
    - OSIS-Konzepte und Methoden
    - OSIS-Spezialisierung
    - OSIS-Techniken und Werkzeuge
  - Data Engineering MA
    - DATA-Konzepte und Methoden
    - DATA-Techniken und Werkzeuge
    - DATA-Spezialisierung
    - SCAL-Konzepte und Methode
    - SCAL-Techniken und Werkzeuge
    - SCAL-Spezialisierung
  - Digital Health MA
    - SCAD-Concepts and Methods
    - SCAD-Technologies and Tools
    - SCAD-Specialization
    - APAD-Concepts and Methods
    - APAD-Technologies and Tools
    - APAD-Specialization

Grading

- Final grade consists of
  - Preparation quiz (mandatory)
  - Group work, presentation, and participation during the block week (40%)
  - Written or oral exam, depending on #participants (60%)

Schedule



Group Assignment – 28th of April

De-Registration Deadline – 12th of May

Quiz Deadline – 2th of June

Group Work & Prep

Preparation Meeting – July or **September**

Start Block Week – One week later

Block Week

End Block Week

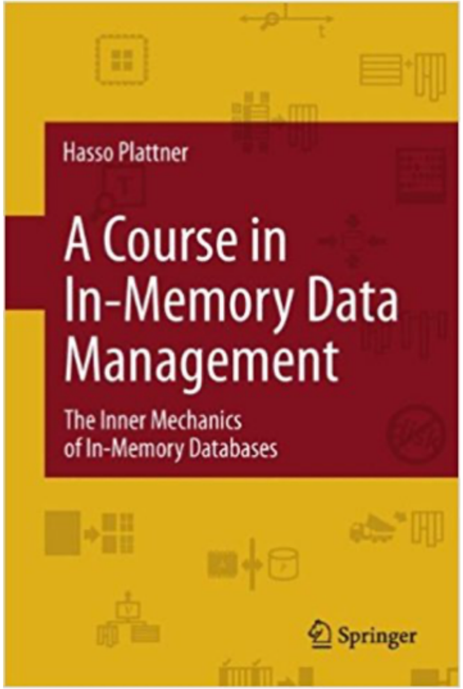
Exam – End of September

Preparation Quiz

- Get a solid understanding of the fundamentals

- Materials
  - □ Course book (distributed digitally)
  - □ openHPI course
    https://open.hpi.de/courses/tuk2020

- Mandatory quiz
  - □ Start: 26th of April
  - □ Deadline: 2nd of June


Hasso Plattner
A Course in In-Memory Data Management
The Inner Mechanics of In-Memory Databases
Springer

Further Readings

Plattner & Leukert:

The In-Memory Revolution

How SAP HANA Enables Business of the Future

THE IN-MEMORY REVOLUTION
HOW SAP HANA ENABLES BUSINESS OF THE FUTURE
Hasso Plattner & Bernd Leukert
Springer

Reznik, Dobson & Gienow:

Cloud Native Transformation:

Practical Patterns for Innovation

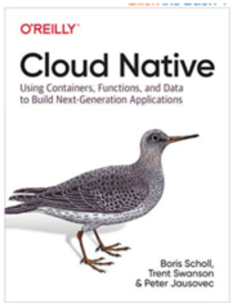
O'REILLY
Cloud Native Transformation
Practical Patterns for Innovation
Pini Reznik, Jamie Dobson & Michelle Gienow

Scholl, Swanson & Jausovec:

Cloud Native:

Using Containers, Functions, and Data

to Build Next-Generation Applications


O'REILLY
Cloud Native
Using Containers, Functions, and Data to Build Next-Generation Applications
Boris Scholl, Trent Swanson & Peter Jausovec

Thematic Structure of Block Week

- Structured along the following overarching topics:
  - **Day 1:** The In-Memory Revolution
    - Architectural Basics and Historic Development of In-Memory Databases
  - **Day 2:** Scaling to Enterprise Needs
    - Replication, Partitioning and Tuning of In-Memory Computing
  - **Day 3:** In-Memory Shifting to the Cloud
    - Requirements, Competitors and Status Quo
  - **Day 4:** Impact on Business
    - Use cases of today and tomorrow

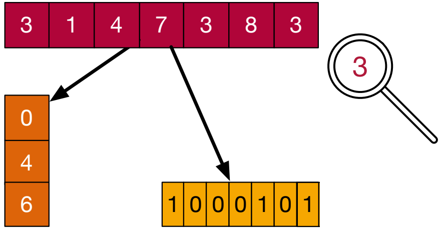
Group Work

- Preparation of interactive group part
  - □ Teams of 6 to 8 students
  - □ Regular meetings
  - □ Team assignment: 28th of April

- Hands-on experiments
  - □ Familiarization with existing research
  - □ Implementation part in C/C++
  - □ Evaluation of the results
  - □ Presentation in the block week (~30 minutes)

- Tell us your topic preference: https://forms.gle/Rhp7TYSRKLQR92Xm7

Topic 1: Many ways to scan a list of integers



Motivation

A significant advantage of in-memory databases is the speed at which dictionary-compressed columns can be scanned. But even the seemingly trivial task of scanning a list of integers holds a surprising number of performance challenges.

Challenges

- Find the most efficient format for writing results
- Stop CPU “features” such as branch mispredictions from affecting your performance
- Do “between” comparisons with a single comparison
- Convince SIMD to work to your advantage
- Beat Hyrise

Learnings

- Low-Level C(++) Programming
- Proper Benchmarking

Requirements

- Basic C++ knowledge (pointers, vectors) is expected

Topic 2: Physical Optimization of Stored Data

Motivation

Reorganizing the stored data in a database allows for various performance improvements in analytical systems. When searching on sorted data, binary searches can be used over linear searches. Moreover, clustering data in clusters of similar data characteristics allows the database to skip large parts of the data without ever looking at it.

Challenges

- Understand the storage layout and architecture of modern in-memory systems
- Understand where sorting and clustering are applicable and in which cases they are not
- Determine on which dimensions to sort/cluster given an actual database workload

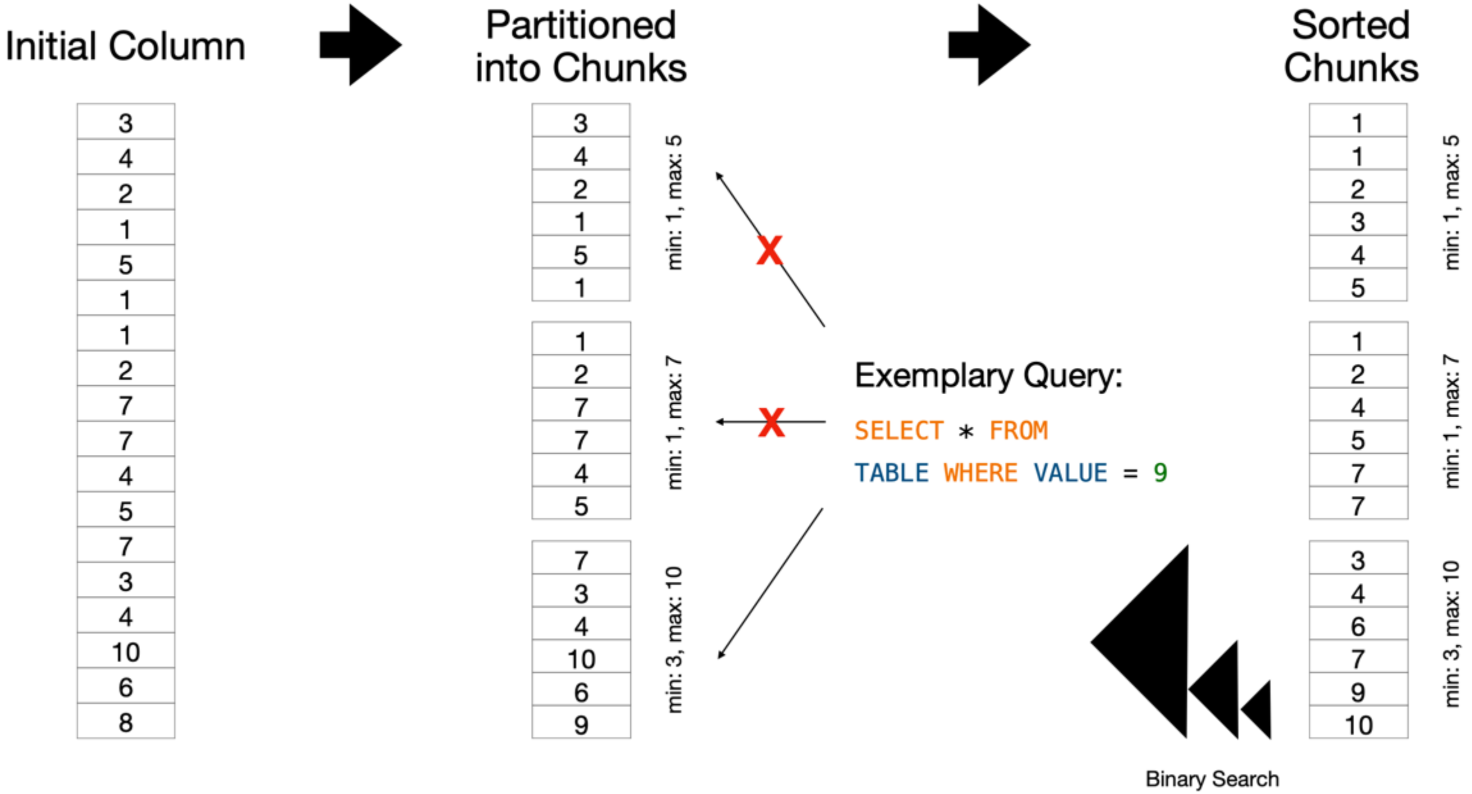
Learnings

- First looks into the in-memory database Hyrise and its storage engine
- Proper Benchmarking

Requirements

- Basic C++ knowledge (pointers, vectors) is advantageous

Topic 2: Physical Optimization of Stored Data



Initial Column	Partitioned into Chunks		Sorted Chunks	
3	3	min: 1, max: 5	1	min: 1, max: 5
4	4		1	
2	2		2	
1	1		3	
5	5		4	
1	1		5	
1	1	min: 1, max: 7	1	min: 1, max: 7
2	2		2	
7	7		4	
7	7		5	
4	4		7	
5	5		7	
7	7	min: 3, max: 10	3	min: 3, max: 10
3	3		4	
4	4		6	
10	10		7	
6	6		9	
8	9		10	

Exemplary Query:

```
SELECT * FROM
TABLE WHERE VALUE = 9
```

Topic 3: Choosing your Database System in the Cloud

Motivation

For any given application workload, there today is a variety of cloud-based database offerings. The underlying database systems are based on different architectures with respective tradeoffs. These need to be understood for an educated choice between the offerings. This year’s focus is on analytics.

Challenges

- Run a representative analytics workload and interpret its results with respect to time and cost
- Tune a database system for a given workload
- Understand the architectures and tradeoffs of current cloud database offerings, i.e, be able to decide when to use what and to explain why

Learnings

- Hands-on experience with modern cloud databases
- Proper benchmarking

Requirements

- Basic database knowledge is expected

Topic 4: Scanning the Cloud in Seconds and Dollars

Motivation

Modern cloud infrastructure separates compute and storage resources in order to provision and scale them independently. So-called *cloud-native* databases build upon this infrastructure: Their query operators run in VMs and their data lies in shared object storage services. But how do they do scans efficiently?

Challenges

- Determine an efficient data format for reading and writing to object stores
- Identify and resolve bottlenecks when dealing with remote cloud storage
- Exploit the available parallelism (within and across single machines)
- Consider performance as well as cost of your solution (and try beat Skyrise)

Learnings

- C++ programming
- Cloud service SDKs and APIs
- Proper benchmarking

Requirements

- Basic C++ knowledge is expected

Contacts

- Michael Perscheid
  - Email: michael.perscheid@hpi.de

- Ralf Teusner
  - Email: ralf.teusner@hpi.de

Tell us your topic preference:
https://forms.gle/Rhp7TYSRKLQR92Xm7