Systems Infrastructure for Data Science

Web Science Group
Uni Freiburg
WS 2012/13

Introduction

About myself

- Since Fall 2011:
  Assistant/Junior Professor for Web Science
- Before Senior Researcher (“Oberassistent”) at Systems Group, ETH Zurich
- Research Interests:
  - Realtime Analytics
  - Stream/Event Processing
  - Social Media Analytics
  - XML/Web Technologies
- Systems-oriented approach

Basic Course Information

- Credits: 3V + 1U (= 6 ECTS)
- Language: English (feel free to ask in German)
- Time and Location:
  - Monday 14:15-16:00 (SR 00-007, Building 106)
  - Fridays 14:15-15:00 (SR 00-007, Building 106)
  - Exercise: Fridays 15:00-15:45 (also SR 00-007)
- Webpage:
  - https://websci.informatik.uni-freiburg.de/teaching/ws201213/infosys/infosys

Workload & Grading

- Exercises
  - Weekly exercise sheets with questions related to the lecture coverage
  - Not graded
  - Attendance to exercise sessions is not mandatory, but it is highly recommended to do well in the exam.
- Exam
  - No prerequisites to participate
  - Written or oral dependent on number of participants

Exercise Sessions

- Content
  - Explain the new exercise sheet
  - Provide solutions for the previous exercise sheet
  - Answer your questions
- First sheets will be made available on October 26th
- The first exercise sessions will take place on November 2nd

Course Objectives

- Overall Objective:
  - to understand how different platforms for data management and analysis work
- Partial goals:
  - Understand the internals of the architecture, implementation, and optimization of a relational database system
  - Understand the basics of distributed databases
  - Understand concepts and implementations of novel platforms

Course Motivation: New Analytics

- No longer just structured, „clean“ business data:
  - Text data, photos, videos
  - Social media: social networks, social streams
  - Science
  - …
- Much broader range of analytics
  - Information Retrieval
  - Machine Learning: Classification, Mining
  - Statistics
  - Human Interaction: Crowdsourcing, Interactive exploration
- Much larger volumes (think Google, Facebook!)
- Unpredictable workloads
- Results required in realtime

Course Motivation: New Platforms

- Increasing CPU core count: Massive Parallelism
- Increasing RAM, „slower“ disks, new storage
- Faster Networks and massive Distribution
  - Racks and Datacenters as new basic building blocks
  - Global Replication, Consistency and Access
- New Processing paradigms:
  - Map/Reduce
  - Key/Value Stores
  - Event, Data Stream Processing

Topics

- Classical Databases: “Complete” package for moderate workloads
  - Storage and Indexing
  - Query Processing and Optimization
  - Performance Tuning and Benchmarking
- Distributed Databases: Scaling with DB means
  - General Architecture
  - Distribution
  - Query Processing
- Map-Reduce: Highly scalable, unstructured data, simple programming model
- Key-Value Stores: Storing and retrieving data
- Stream Processing: Processing instantly without storing

Relation to other lectures

- Information Systems:
  - DB Intro: Foundations, Transactions
  - Distributed Systems: focused on data consistency and distributed transactions
  - Data Models and Query Languages: covers models, languages and theory
- Other areas:
  - Operating Systems, Networks: same foundations, sometimes same problems

Starting point: Classical DB

- Still useful for moderate-sized workloads (few TBs, standard queries, OLTP)
- Guidepost for techologies
  - Nearly all aspects of data management covered
  - Decades of experience and refinement
  - (Many aspects being re-discovered by „cool new platforms“)

Architecture of a Database System

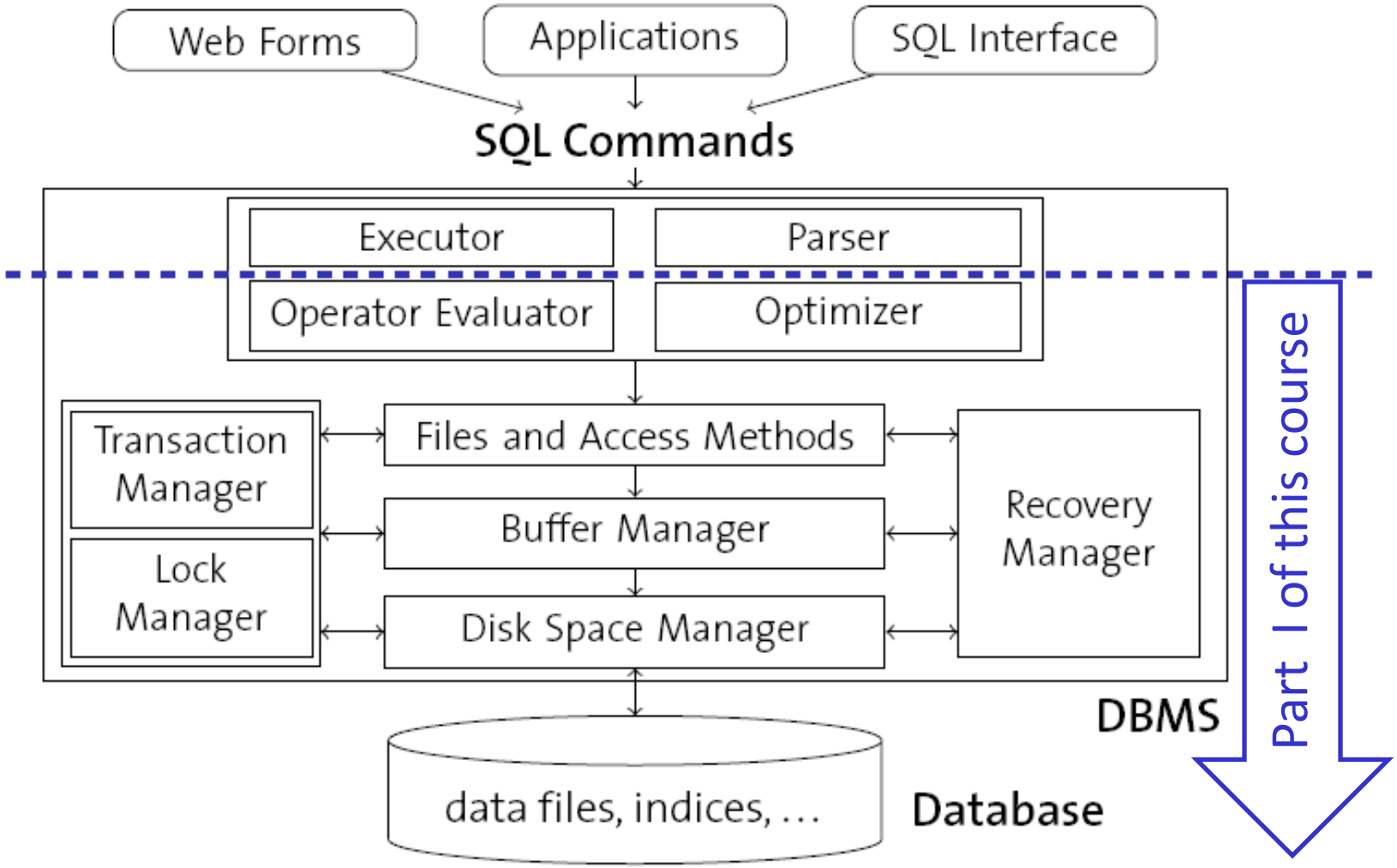

Figure inspired by Ramakrishnan/Gehrke: "Database Management Systems", McGraw-Hill 2003.