CSE 444: Database Internals

Lecture 1
Introduction

Staff

- Instructor: Magdalena Balazinska (magda)
  - magda@cs.washington.edu
    Office hours: Mondays 10am-12pm in CSE 584
- TA: Shengliang Xu
  - sxlu@cs.washington.edu
  - Office hour: Tuesday 1pm-2pm in CSE 220
- 1/2 TA: Prasang Upadhyaya
  - prasang@cs.washington.edu
  - Office hour: Friday 2:30pm-3:30pm in CSE 2XX

About Me: General

- At UW since January 2006
- PhD from MIT
- Born in Poland
- Grew-up in Poland, Algeria, and Canada

About Me: Research

- Past: Stream Processing
  - Distributed stream processing (Borealis)
  - RFID data management (RFID Ecosystem)
  - Probabilistic event processing (Lahar)

- Now: Cloud computing and scientific data mgmt
  - Collaboration with astronomers, oceanographers, etc.
  - Making large-scale data analysis easier and faster
  - Helping users leverage cloud computing
  - Interactions between pricing and data management

Course Goals

- The world is drowning in data!
- Need computer scientists to help manage this data
  - Help domain scientists achieve new discoveries
  - Help companies provide better services (e.g. Facebook)
  - Help governments become more efficient
- This class: **principles of building data mgmt systems**
  - Learn how classical DBMSs are built
  - Learn key principles and techniques
  - Get hands-on experience building a (parallel) DBMS

Course Format

- Lectures MWF, 12:30pm-1:20pm
- Sections: Th 9:30-10:20, 10:30-11:20
  - Content: exercises, tutorials, questions
  - Location: See course website

- 6 Homeworks: programming assignments (labs)
- 6 Short web quizzes

- Midterm and final

Communications

- Web page: http://www.cs.washington.edu/444
  - Lectures will be available there (see calendar)
  - Homeworks will be available there
  - Web quizzes will be available there
- Mailing list
  - Announcements, group discussions
  - You are already subscribed
- Message board
  - Great place to ask assignment-related questions

Textbook

Main textbook, available at the bookstore:

- *Database Systems: The Complete Book*, Hector Garcia-Molina, Jeffrey Ullman, and Jennifer Widom
  Second edition.

See course website for recommended chapters

Most important: COME TO CLASS ! ASK QUESTIONS !

Other Readings

- See course website (Calendar.html) for extra readings for some of the lectures: e.g. MapReduce paper.

- Other books available at the Engineering Library (not on reserve):
  - *Database Management Systems*, Ramakrishnan & Gerhke
  - *Fundamentals of Database Systems*, Elmasri, Navathe
  - *Foundations of Databases*, Abiteboul, Hull, Vianu

Grading

- Homeworks/Labs   45%
- Web quizzes 10%
- Midterm   20%
- Final   25%

Six Homeworks/Labs

- Lab 1: Build a DBMS that can scan a relation on disk
  - **Part 1 of this lab is due next Monday!**
- Lab 2: Build a DBMS that can run simple SQL queries and also supports data updates
  - There will be a contest for the fastest DBMS
- Lab 4: Add a query optimizer
- Lab 3: Add a lock manager (transactions)
- Lab 5: Add a write-ahead log (transactions)
- Lab 6: Make your DBMS parallel!

Warning: I **will** run cheating-detecting software!

About the Homeworks/Labs

Logistics:
- To be done **INDIVIDUALLY!**
- Each lab will take a **significant** amount of time
- Labs build on each other
Purpose
- Hands-on experience building a DBMS
- Deepen your understanding significantly
- We will build a *classical* DBMS
- After graduation, you will build *new-types* of DBMSs
  - Will know the principles and will make your own design choices

Six Web Quizzes

- Short online tests. **First one due next week!**
- Can take many times: best score counts!
- Provide explanations for wrong answers
- Will help you
  - Test your knowledge
  - Stay in synch with lectures
- Most difficult: remember to do them!

Check course website for due dates
Access token is on the board

13

Exams

- Midterm and Final

- Check course website for dates
- Location: in class

- I will post practice problems since 444 has been re-designed

Outline (this lecture and next)

- Review of DBMS goals and features

- Review of relational model

- Review of SQL

DBMS Goals

What is a database management system (DBMS)?
- A DBMS is a software system designed to provide data management service

Why do DBMSs exist?
- Data is valuable
- Managing that data is difficult
  - See next slide for some of the required features
- A DBMS simplifies data management task
- A DBMS facilitates application development

DBMS Key Features

How to decide what features should go into the DBMS?

- Data independence
  - Data model
  - Data definition language
  - Data manipulation language
- Efficient data access
- Data integrity and security
- Data administration
- Concurrency control
- Crash recovery

Key Data Mngmt Concepts

- Data models: how to describe real-world data
  - Relational, XML, graph data (RDF)
- Schema vs data
- Declarative query languages
  - Say what you want not how to get it
- Data independence
  - Physical: Can change how data is stored on disk without maintenance to applications
- Query compiler and optimizer
- Transactions: isolation and atomicity

Course Content

Focus: how to build a classical relational DBMS
- Review of the relational model (lecture 1 and 2)
- DBMS architecture and deployments (lecture 3)
- Data storage, indexing, and buffer mgmt (lectures 4-6)
- Query evaluation (lectures 7-9)
- Query optimization (lectures 10-13)
- Transactions (lectures 14-19)
- Replication and distribution (lectures 20-22)
- Parallel query processing (lectures 22-25)
- Database as a service and NoSQL (lectures 26 and 27)

Relational Model...

- Let's start our review of the relational model…
- We will continue next lecture