CSE 444: Database Internals

Lecture 1
Introduction

Course Instructors

- Magdalena Balazinska (Magda)
  - magda@cs.washington.edu
    Office hour: Tuesday 4:30pm-5:30pm in CSE 584
- Sudeepa Roy (Sudeepa)
  - sudeepa@cs.washington.edu
  - Office hour: Wednesday, 4:30pm-5:30pm, CSE 344

Course TAs

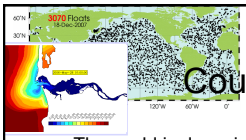
- TA: Jingjing Wang
  - jingjing@cs.washington.edu
  - Office hour: Monday, 4:30pm-5:30pm, CSE 220
- ½ TA: Shengliang Xu
  - sxlu@cs.washington.edu
  - Office hour: Friday, 4:30pm-5:30pm, CSE 220

About Me: General

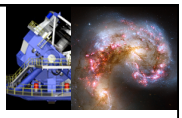
- At UW since January 2006
- PhD from MIT
- Born in Poland
- Grew-up in Poland, Algeria, and Canada

About Me: Research

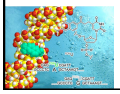
- Past: Stream Processing
  - Distributed stream processing (Borealis)
  - RFID data management (RFID Ecosystem)
  - Probabilistic event processing (Lahar)
- Now: Cloud computing and scientific data mgmt
  - Collaboration with astronomers, oceanographers, etc.
  - Making large-scale data analysis easier 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
- Homeworks
  - 6 Labs: programming assignments
  - 6 Homeworks: written assignments
- NO exams: This is a project-oriented course
  - In this course you will learn by doing… so expect to spend a lot of time on the "doing"

Communication (part 1)

- Web page: http://www.cs.washington.edu/444
  - Lectures will be available there (see calendar)
  - Homeworks/Labs will be available there
- Mailing list
  - Announcements, group discussions
  - You are already subscribed

Communication (part 2)

- Message board
  - **Place to ask questions about the course, labs, homeworks, etc.**
  - **Remember NOT to post any fragment of your code**
  - Do NOT send questions by email to the course staff unless
    - You need to discuss a personal matter
    - You want to setup an appointment
    - A question has not been answered on the message board

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 highly recommended book:
  - *Database Management Systems*, Ramakrishnan & Gerhke
  - Use this book if you do not like the Ullman book

Grading

- Lab 1, 2, 3, and 5: 40% (10% each)
- Final Lab 4 or Lab 6 (your choice): 15%
- Final project report 10%
- Six written assignments: 35%

Six Labs

- Lab 1: Build a DBMS that can scan a relation on disk
  - **Part 1 of this lab is due on Friday!**
- 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 3: Add a lock manager (transactions)
- Lab 5: Add a write-ahead log (transactions)
- Lab 4: Add a query optimizer
- Lab 6: Make your DBMS parallel!

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

About the 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 Homeworks

- Written assignments
- Help review material learned in class
- Prepare you for the labs
  - One homework before each corresponding lab
- Go beyond what we implement in labs
- To be done **INDIVIDUALLY**

Exams

- No exams

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

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

How to decide what features should go into the DBMS?

Some Key Data Management 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)
- Parallel query processing (lectures 20-22)
- Replication and distribution (lectures 23-25)
- 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