Introduction to Data Management
CSE 344

Lecture 1: Introduction

Couldn’t register?
Signup on overload list http://tinyurl.com/hz9sxzd
Ask me for the code word after class

Webquiz token (write this down):
37FE0390
Class 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
• Welcome to 344: Introduction to Data Management
  – Existing tools PLUS data management principles
  – This is not just a class on SQL!
Turing Awards in Data Management

Charles Bachman, 1973
*IDS and CODASYL*

Ted Codd, 1981
*Relational model*

Jim Gray, 1998
*Transaction processing*

Michael Stonebraker, 2014
*INGRES and Postgres*

You could be next!!
Staff

• Instructor: Alvin Cheung
  – Office hour on Thursdays, 11am-noon in CSE 530

From ACM Spring BBQ 15
Staff

• **TAs:**
  - Nicholas Anderson
  - Shumo Chu
  - Kelly Jiang
  - Clara Lu
  - Jonathan Phippen
  - Amarpal Singh
  - Cindy Suripto
  - Lisa Zhang

• See course website for office hours and locations

• **Contacting staff:**
  - Please use piazza and anonymous feedback link on course website
  - All course announcements will be posted on piazza, make sure you sign up
Course Format

• Lectures MWTh, 3:30-4:20 pm
  – Location: here!

• Sections: Thursdays
  – Content: exercises, tutorials, questions
  – Locations: see web
  – We will take attendance

• 7 homework assignments
• 6 web quizzes

• Midterm and final

• Class and section participation
  – Post and answer questions (in class, piazza, etc)
  – In-class exercises (hint: come to class!)
Grading

• Homeworks 30%
• Web quizzes 10%
• Midterm 20%
• Final 30%
• Class participation 10%

• This is all subject to change
Communications

• **Web page:** [http://www.cs.washington.edu/344](http://www.cs.washington.edu/344)
  - Syllabus is there
  - Lectures will be available there (see calendar)
  - Homework assignments will be available there
  - Link to web quizzes is there

• **Piazza**
  - Make sure you sign up:
  - **THE** place to ask course-related questions
  - Log in today and enable notifications
Textbook

Main textbook, available at the bookstore:


**Second edition.**

Textbook (and others) are REQUIRED READING!

Most important: COME TO CLASS! ASK QUESTIONS!
Other Texts

Available at the Engineering Library
(some on reserve):

• *Database Management Systems*, Ramakrishnan
• *Fundamentals of Database Systems*, Elmasri, Navathe
• *Foundations of Databases*, Abiteboul, Hull, Vianu
• *Data on the Web*, Abiteboul, Buneman, Suciu
Seven Homework Assignments

H1&H2: Basic SQL with SQLite
H3: Advanced SQL with SQL Server
H4: Relational algebra, Datalog
H5: NoSQL
H6: Conceptual Design
H7: SQL in Java (JDBC)

Check calendar for due dates -- Submit via dropbox!
About the Assignments

• Homework assignments will take time but most time should be spent *learning*

• Do them on your own

• Very practical assignments

• Put everything on your resume!!!
  – SQL, SQLite, SQL Server, SQL Azure JDBC, JSON, Hadoop,…
Deadlines and Late Days

• Assignments are expected to be done on time, but things happen, so…
• You have up to 4 late days
  – No more than 2 on any one assignment
  – Use in 24-hour chunks
• Late days = safety net, not convenience!
  – You should not plan on using them
  – If you use all 4 you are doing it wrong
Six Web Quizzes

- [http://newgradiance.com/](http://newgradiance.com/)
- Create account, provide token
- **Class token: 37FE0390** (will post on piazza)
- Short tests, take many times, best score counts
- **No late days** – closes at 11:00 deadline
- Provide explanations for wrong answers
- Will help you
  - Test your knowledge
  - Stay in synch with class
  - Get ready for homework assignments
Exams

• Midterm and Final
  – See course calendar for dates and times

• Can bring letter-size piece of paper with notes
  – Can write on both sides
  – Midterm: 1 sheet, Final: 2 sheets

• Closed book. No computers, phones, watches, etc.!

• Check course website for dates

• Location: in class
Academic Integrity

• Anything you submit for credit is expected to be your own work
  – Of course OK to exchange ideas, but not detailed solutions
  – We all know difference between collaboration and cheating
  – Attempt to gain credit for work you did not do is misconduct

• I trust you implicitly, but will come down hard on any violations of that trust
Lecture Notes

• Will be available before class online
• Feel free to bring them to class to take notes
Using Electronics in Class

• Opened laptops create disturbances to your neighbors
• Please sit in the back if you use your laptop to take notes
• OK if you use surfaces

• And please don’t check your email / sms / youtube / fb / etc during class
  – If people are doing this we will have to ban all laptops 😞
Student evals from fall 16…

- All the class materials are very interesting and mind-opening… Like it a lot!
- The practical applications of the concepts we learned such as sql queries, java application programming, and using cloud storage were all very informative.
- The instructor's enthusiasm. First half of class seemed very well designed.
- Alvin's humor and explanations of many difficult concept
Student evals from fall 16...

- More clarifications on homework setup.
- Faster grading
- Slow down a bit when teaching during class.
- The instructor's voice was often mumbled or quiet during lectures.
- Nothing. Thanks for the quarter! I felt I learned a lot.

Lesson: Do not wait till course evals to suggest changes!
Now onto the real stuff...
Outline of Today’s Lecture

• Overview of database management systems
  – Why they are helpful
  – What are some of their key features
  – What are some of their key concepts

• Course content
Database

What is a database?
Database

What is a database?
• A collection of files storing related data

Give examples of databases
Database Management System

What is a DBMS?

Give examples of DBMSs
Database Management System

What is a DBMS?

• A big program written by someone else that allows us to manage efficiently a large database and allows it to persist over long periods of time

Give examples of DBMSs

– Oracle, IBM DB2, Microsoft SQL Server, Vertica, Teradata
– Open source: MySQL (Sun/Oracle), PostgreSQL, CouchDB
– Open source library: SQLite

We will focus on relational DBMSs most quarter
An Example: Online Bookseller

• What data do we need?
  –
  –
  –
  –

• What capabilities on the data do we need?
  –
  –
  –
An Example: Online Bookseller

• What data do we need?
  – Data about books, customers, pending orders, order histories, trends, preferences, etc.
  – Data about sessions (clicks, pages, searches)
  – Note: data must be persistent! Outlive application
  – Also note that data is large… won’t fit all in memory

• What capabilities on the data do we need?
An Example: Online Bookseller

• What data do we need?
  – Data about books, customers, pending orders, order histories, trends, preferences, etc.
  – Data about sessions (clicks, pages, searches)
  – Note: data must be persistent! Outlive application
  – Also note that data is large… won’t fit all in memory

• What capabilities on the data do we need?
  – Insert/remove books, find books by author/title/etc., analyze past order history, recommend books, …
  – Data must be accessed efficiently, by many users
  – Data must be safe from failures and malicious users
Multi-user discussion

• Jane and John both have ID number for gift certificate (credit) of $200 they got as a wedding gift
  – Jane @ her office orders "The Selfish Gene, R. Dawkins" ($80)
  – John @ his office orders "Guns and Steel, J. Diamond" ($100)

• Questions:
  – What is the ending credit?
  – What if second book costs $130?
  – What if system crashes?
Discussion

• Did you ever encounter a data management problem?
  – Experimental data from a homework?
  – Personal data?
  – Other data?

• How did you manage your data?
Summary Required Data Management Functionality

1. Describe real-world entities in terms of stored data
2. Persistently store large datasets
3. Efficiently query & update
   - Must handle complex questions about data
   - Must handle sophisticated updates
   - Performance matters
4. Change structure (e.g., add attributes)
5. Concurrency control: enable simultaneous updates
6. Crash recovery
7. Security and integrity
DBMS Benefits

• Expensive to implement all these features inside the application

• DBMS provides these features (and more)

• DBMS simplifies application development
Who are the players?

• **DB application developer**: writes programs that query and modify data (344)
• **DB designer**: establishes schema (344)
• **DB administrator**: loads data, tunes system, keeps whole thing running (344, 444)
• **Data analyst**: data mining, data integration (344, 446)
• **DBMS implementor**: builds the DBMS (444)
Key Data Management Concepts

• **Data models**: how to describe real-world data
  – Relational, NoSQL, …

• **Declarative query language**
  – Say what you want not how to get it

• **Data independence**
  – Physical independence: Can change how data is stored on disk without maintenance to applications
  – Logical independence: can change schema w/o affecting apps

• **Query optimizer** and compiler

• **Physical design**

• **Transactions**: isolation and atomicity
What is this class about?

- **Focus: Using DBMSs**
- Relational Data Model
  - SQL, Relational Algebra, Relational Calculus, datalog
- Semistructured Data Model
  - JSON (NoSQL)
- Conceptual design
  - E/R diagrams, Views, and Database normalization
- Transactions
- Parallel databases and MapReduce
What to Do Now

http://www.cs.washington.edu/344

• Webquiz 1 is open
  – Create account at http://newgradiance.com/
  – Sign up for class online
  – Due on Tuesday 1/10, 11 pm

• Homework 1 is posted
  – Simple queries in SQL Lite
  – Due on Wednesday 1/11, 11 pm

• Sections tomorrow
  – Tutorial on SQL Lite

• Lecture tomorrow
  – Data models

• Sign up on overload website if you’re still trying to register
• Post on Piazza if you have questions about HW and lecture