COURSE FORMAT

Lectures
• Location: SIG 134
• Please attend

Sections:
• Content: exercises, tutorials, questions, new materials (occasionally)
• Locations: see web
• Please attend
• Bring your laptop

8 homework assignments
7 web quizzes

Midterm and final
GRADING

Homeworks 30%
Web quizzes 10%
Midterm 25%
Final 35%

This is all subject to change
COMMUNICATIONS

Web page: http://www.cs.washington.edu/344
- Syllabus (course information)
- Lecture/section notes will be available there
- Homework assignments will be available there
- Link to web quizzes is there

Piazza
- Sign up: Link Soon
- THE place to ask course-related questions
- Log in today and enable notifications

Class mailing list
- You are automatically subscribed
- Low traffic, only important announcements
TEXTBOOK

Main textbook, available at the bookstore:


Second edition.
EIGHT HOMEWORK ASSIGNMENTS

H1: Sqlite intro (1 wk)
H2: Sqlite basics (1 wk)
H3: Advanced SQL on Azure (1⅔ weeks)
H4: Datalog and Relational Algebra (1⅓ weeks)
H5: NoSQL: Json/SQL++ (1 wk)
H6: Spark on AWS (1⅓ weeks)
H7: Schema Design (1wk)
H8: Transactional Application (1⅓ weeks)

New this year: submit via git
ABOUT THE ASSIGNMENTS

You will learn/practice the course material:

• SQL, RA, parallel db, transactions, ...

You will also learn lots of new technology

• Cloud computing: Azure, Cloud9, AWS
• NoSQL: AsterixDB, LogicBlox
• Git

The time spent learning the new technology is very useful: write everything on your CV!
DEADLINES AND LATE DAYS

Assignments are expected to be done on time, but things happen, so...

You have up to 3 late days

- No more than 2 on any one assignment
- Used in 24-hour chunks

Late days = safety net, not convenience!

- You should not plan on using them
- If you use all 3 you are doing it wrong
SEVEN WEB QUIZZES

- http://newgradiance.com/
- Create account; please use the same ID as your UW ID
- Token to be provided to course email

Short tests, take many times, best score counts

No late days – closes at 11:00 deadline

Provide explanations for wrong answers
EXAMS

Midterm (TBA – Early February)
Final, Thursday, March 15th, 230-4:20

Closed book. No computers, phones, watches,...

Location: in class
ABOUT ME

• Evan McCarty (ejmcc@cs.washington.edu)
• Theory and Algorithms research
• Data Scientist for Partners for Our Children
• Lecture notes posted after class
  • Panopto recordings
• Part-time Faculty
  • On campus MWF
  • Available by email
• Office hours
  • Monday and Friday 4:30 – 6:00 or by email.
ABOUT STAFF

• TAs
  • Joshua Bean
  • Allison Chou
  • Colin Evans
  • Jayanth Garlapati
  • Jonathan Leang
  • Cindy Suripto
  • James Wang

• First resource for coding / setup problems
ABOUT YOU

• Expect most are CSE majors

• (Hopefully) registered
  • If not https://docs.google.com/forms/d/e/1FAIpQLSf4hqZmELivR1_lby_WmpgT66OM78K-Ed-suebQTI84B0SLow/viewform

• Academic Honesty and Participation

• Piazza and help
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 (and universities!) become more efficient

Welcome to 344: Introduction to Data Management

• Existing tools PLUS data management principles
• This is not just a class on SQL!
WHY DATABASE MANAGEMENT?

- This course was my least favorite topic in undergrad
WHY DATABASE MANAGEMENT?

• This course was my least favorite topic in undergrad
• Now, I work with databases
WHY DATABASE MANAGEMENT?

• This course was my least favorite topic in undergrad
• Now, I work with databases
  • Intelligent design and organization of data allows important work and research to occur efficiently and correctly
WHY DATABASE MANAGEMENT?

• This course was my least favorite topic in undergrad
• Now, I work with databases
  • Intelligent design and organization of data allows important work and research to occur efficiently and correctly
• Organizations need a diverse set of skills, you may not ever need to manage a DB, but you will certainly be interfacing with one
WHY DATABASE MANAGEMENT?

• This course was my least favorite topic in undergrad

• Now, I work with databases
  • Intelligent design and organization of data allows important work and research to occur efficiently and correctly

• Organizations need a diverse set of skills, you may not ever need to manage a DB, but you will certainly be interfacing with one

• Decisions made in setting up a DB (or even a query) can affect performance going forward
WHY DATABASE MANAGEMENT?

• Disk and magnetic tape are linear storage
  • We can access elements throughout them, but there is a continuous serialization of this data.
  • Data itself is rarely one dimensional
  • Imagine storing all data about UW students on disk
WHY DATABASE MANAGEMENT?

- Disk and magnetic tape are linear storage
  - We can access elements throughout them, but there is a continuous serialization of this data.
  - Data itself is rarely one dimensional
  - Imagine storing all data about UW students on disk
- What is their order? Are students related?
WHY DATABASE MANAGEMENT?

• Disk and magnetic tape are linear storage
  • We can access elements throughout them, but there is a continuous serialization of this data.
  • Data itself is rarely one dimensional
  • Imagine storing all data about UW students on disk
• What is their order? Are students related?
  • Related relative to other data?
  • Why store “students” at all?
What is a database?
DATABASE

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

Give examples of databases
What is a database?
A collection of files storing related data.

Give examples of databases
Accounts database; payroll database; UW’s students database; Amazon’s products database; airline reservation database.
DATABASE MANAGEMENT SYSTEM

What is a DBMS?
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.

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
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
CHALLENGES FOR A DBMS

Alice and Bob receive a $200 gift certificate as wedding gift
CHALLENGES FOR A DBMS

Alice and Bob receive a $200 gift certificate as wedding gift

Alice @ her office orders “The Selfish Gene”

Bob @ home orders “Guns, germs, and steel”
Alice and Bob receive a $200 gift certificate as wedding gift

Alice @ her office orders "The Selfish Gene" for $80
Bob @ home orders "Guns, germs, and steel" for $100

Questions:
What is the ending credit?
What if second book costs $130?
What if system crashes?
CHALLENGES FOR A DBMS

Alice and Bob receive a $200 gift certificate as wedding gift

Alice @ her office orders "The Selfish Gene" $80
Bob @ home orders “Guns, germs, and steel” $100

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

Lesson: a DBMS needs to handle various scenarios
WHAT A DBMS DOES

Describe real-world entities in terms of stored data
Persistently store large datasets
Efficiently query & update
  • Must handle complex questions about data
  • Must handle sophisticated updates
  • Performance matters
Change structure (e.g., add attributes)
Concurrency control: enable simultaneous updates
Crash recovery
Security and integrity
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)
WHAT IS THIS CLASS ABOUT?

Unit 1: Intro (today)
Unit 2: Relational Data Models and Query Languages
Unit 3: Non-relational data
Unit 4: RDMBS internals and query optimization
Unit 5: Parallel query processing
Unit 6: DBMS usability, conceptual design
Unit 7: Transactions
Unit 8: Advanced topics (time permitting)
WHAT TO EXPECT SOON

• Course Website
• Syllabus
• Git tutorial / help
• The first HW assignment
• Piazza page
• Canvas page
• Link for online quizzes