Introduction to Data Management
CSE 414

Lecture 1: Introduction

Write down Webquiz token
Class Goals

• The world is drowning in data!
• Need to help manage this data
  – Domain scientists achieve new discoveries
  – Companies provide better services (e.g., Facebook)
  – Governments (and universities!) become more efficient
• Welcome to 414: Introduction to Data Management
  – Existing tools PLUS data management principles
  – This is not just a class on SQL
Staff

Instructor:
• Dan Suciu

TA’s
• Joshua Bean
• Shumo Chu
• Kartik Arora
• Ruta Dhaneshwar
• Pranay Mundra
• Vineeth Varghese
• Ying Wang
• Yihang Wu
Course Format

• Lectures: this room, please attend!
• Sections: for locations, see web; **bring your laptop**

• 8 homework assignments
• 7 web quizzes

• Midterm and final

• Class and section participation:
  Post and **answer** questions (in class, piazza, etc)
Grading

- Homeworks: 30%
- Web quizzes: 10%
- Midterm: 20%
- Final: 30%
- Class participation: 10%
- Extra credit:
  - Some hw have extra credit questions
  - Large # of good answers on piazza

This is all subject to change.
Communications

  - Everything is here

- **Piazza**
  - **THE** place to ask course-related questions
  - Log in today, enable notifications
  - Warning: [canvas.ucdavis.edu/courses/192458/pages/piazza-warning](canvas.ucdavis.edu/courses/192458/pages/piazza-warning)

- **Class mailing list**
  - Very low traffic, only important announcements
Textbook

Main textbook, available at the bookstore or pdf:


*Second edition.*
Eight Homework Assignments

1. Sqlite intro (due this Friday!!)
2. Sqlite basics
3. SQLAzure
4. Datalog and Relational Algebra
5. Json/SQL++
6. Spark
7. Schema Design, JDBC App
8. JDBC App w/ transactions

Submit via gitlab
About the Assignments

• You will learn/practice the course material
• You will also learn lots of new technology

• Note: some familiarity with programming languages and tools is needed
Deadlines and Late Days

• 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
• Absolutely no exceptions after late days exhausted
Seven Web Quizzes

- [http://newgradiance.com/](http://newgradiance.com/)
- Create account
- Please use the same Last-name/ID as for UW
- Provide token (on the whiteboard)
- Short tests, you may take them many times, best score counts
- **No late days** – closes at 11:00 deadline
Exams

- Midterm (May 3rd) and Final (June 10th)

- You may bring letter-size piece of paper with notes
  - Handwritten
  - May write on both sides
  - Midterm: 1 sheet, Final: 2 sheets

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

- Location: in class
Academic Integrity

• Anything you submit for credit is expected to be your own work
  – OK to exchange ideas, not detailed solutions
  – We all know difference between collaboration and cheating

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

• Lecture notes: Website
  – Feel free to bring them to class to take notes
  – Refresh often, since I improve them continuously

• Panopto recordings: canvas
Now onto the real stuff…
Outline of Today’s Lecture

• Overview of database management systems

• Course content
Database

What is a database?

Give examples of databases
Database

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

Give examples of databases
Database

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?

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
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" for $80

Bob @ home orders "Guns, germs, and steel" for $100
Challenges for a DBMS

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" 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?

Lesson: a DBMS needs to handle various scenarios
What a DBMS Does

- Describe real-world entities
- Store large datasets persistently
- Query & update efficiently
- Change structure (e.g., add attributes)
- Handle concurrent updates
- Crash recovery
- Security and integrity
Key Players

- **DB application developer**: writes programs that query and modify data (414)
- **DB designer**: establishes schema (414)
- **DB administrator**: loads data, tunes system, keeps whole thing running (414, 444)
- **Data analyst**: data mining, data integration (414, 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
What to Do Now

http://www.cs.washington.edu/414

• Webquiz 1 is open
  – Create account at http://newgradiance.com/
  – Sign up for class online
  – Due Saturday, 4/6

• Homework 1 is posted
  – Simple queries in SQL Lite
  – Due on Friday, 4/5

• First sections
  – Tutorial on git, and on SQL Lite

• Log in piazza today, enable notifications