Introduction to Data Management
CSE 344

Lecture 1: Introduction

Write down Webquiz token
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!
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*
Most Enterprise AI Models Are Based on Relational Data*

- Retail: > 86% relational
- Insurance: > 83% relational
- Marketing: > 81% relational
- Financial: > 77% relational

* based on 2017 Kaggle survey of 16,000 ML practitioners
* some of us helped bring the relational model to databases
Staff

• Instructor: Dan Suciu
• TA’s
  – Walter Cai
  – Natalie Fetsch
  – Shana Hutchison
  – Jack Khuu
  – Jonathan Leang
  – Ying Wang
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** [piazza.com/washington/fall2019/cse344](http://piazza.com/washington/fall2019/cse344)
  - **THE** place to ask course-related questions
  - Log in today, enable notifications

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

Main textbook, available at the bookstore or pdf:


Second edition.

REQUIRED READING!
Eight Homework Assignments

1. Sqlite intro (1 wk)
2. Sqlite basics (1 wk)
3. SQLAzure (1\(\frac{2}{3}\) weeks)
4. Datalog and Relational Algebra (1\(\frac{1}{3}\) weeks)
5. Json/SQL++ (1 wk)
6. Spark (2 weeks)
7. Schema Design (1 wk)
8. Transactional Application (1\(\frac{2}{3}\) weeks)

Submit via gitlab
About the Assignments

• You will learn/practice the course material
• You will also learn lots of new technology
  – Time spent learning it is useful! Write on your CV!

• Note: we use anonymized assignments to improve the grading tool; to opt out, send me email after the final grade
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
  – You should not plan on using them
  – If you use all 4 you are doing it wrong
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 (Feb 11) and Final (March 18)

• 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

• Lectures 1/14, 1/16 CANCELED

• Makeup: 1/10, 4:30pm, 1/22, 5:30pm
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"

Bob @ home orders
“Guns, germs, and steel”
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”

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"

Bob @ home orders
“Guns, germs, and steel”

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 (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 Do Now

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

• Webquiz 1 is open
  – Create account at http://newgradiance.com/
  – Sign up for class online
  – Due Saturday, 1/12

• Homework 1 is posted
  – Simple queries in SQL Lite
  – Due on Tuesday, 1/15

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

• Log in piazza today, enable notifications