EECS3421 – Introduction to Database Management Systems

Thanks to John Mylopoulos and Ryan Johnson for material in these slides

Overview

- What is a database?
- Course administrivia
- The relational model

What is a database *system*?

- Database: a large, integrated collection of data
- Models *relevant aspects of reality*
  - Entities (teams, players)
  - Relationships (Lionel Messi *plays for* Barcelona)
  - Constraints (*at least* one goalkeeper per team)
  - More recently, active components ("business logic")

- *Database Management System (DBMS):* a software system designed to store, manage, and facilitate access to databases

In the beginning…

- There was *The Mainframe*
  - Cost: millions
  - Watts: millions
  - Size: 2000 m^2
  - MIPS: 0.04
  - Memory: 2kB
  - Storage: 3.5MB (tape)

SAGE (1954)

Few organizations could afford two!

Early computing challenges

- Bare hardware
  - No OS
  - No device drivers
  - No file system

UNIVAC (1951)

=> File Management System

- Time sharing
  - ~100 terminals per mainframe
  - Users share hardware
  - Want to share *data*, too

SABRE (1960)

=> “The Database”

“The Database”

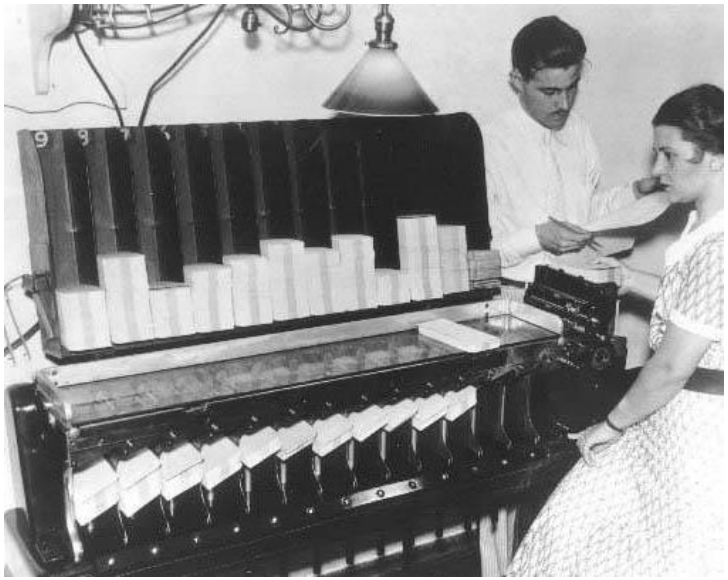
- Abstract concept dating back to the 1950’s
  - *Centralized* repository for all the enterprise’s data
  - *Real-time* updates from many sources
  - *Concurrent* access by many users
  - *Interactive* (ad-hoc) exploration and reporting

Example System: Semi Automatic Ground Environment (SAGE)

- **Goal**: Produce a single unified image of the airspace over an area
- Computer-aided tracking and interception of aircraft
- Dozens of SAGE installations (big one in North Bay)
- Hundreds of radar stations throughout North America
- Thousands of operators

Goal: all relevant information at your fingertips

File management systems (FMS)



- File management ca. 1935
  - **File**: box of punchcards
  - **Metadata**: label on the box
  - **Ad-hoc report**: no big deal
  - **Hardware change**: no big deal
- File management ca. 1955
  - **File**: several km of magnetic tape
  - **Metadata**: embedded in application logic
  - **Ad-hoc report**: hire a couple programmers
  - **Hardware change**: hire a dozen programmers…

Huge need for portability, abstraction

Database Management System

- File management systems meet *The Database*
  - Protect users from each other (isolation, consistency)
  - Protect application from data changes (at logical level)
  - Protect data from hardware changes (at physical level)
- Split personality remains to this day
  - Theory/applications (declarative access to changing data)
  - Systems (make it run fast on ever-changing hardware)

This semester: the theory/application side

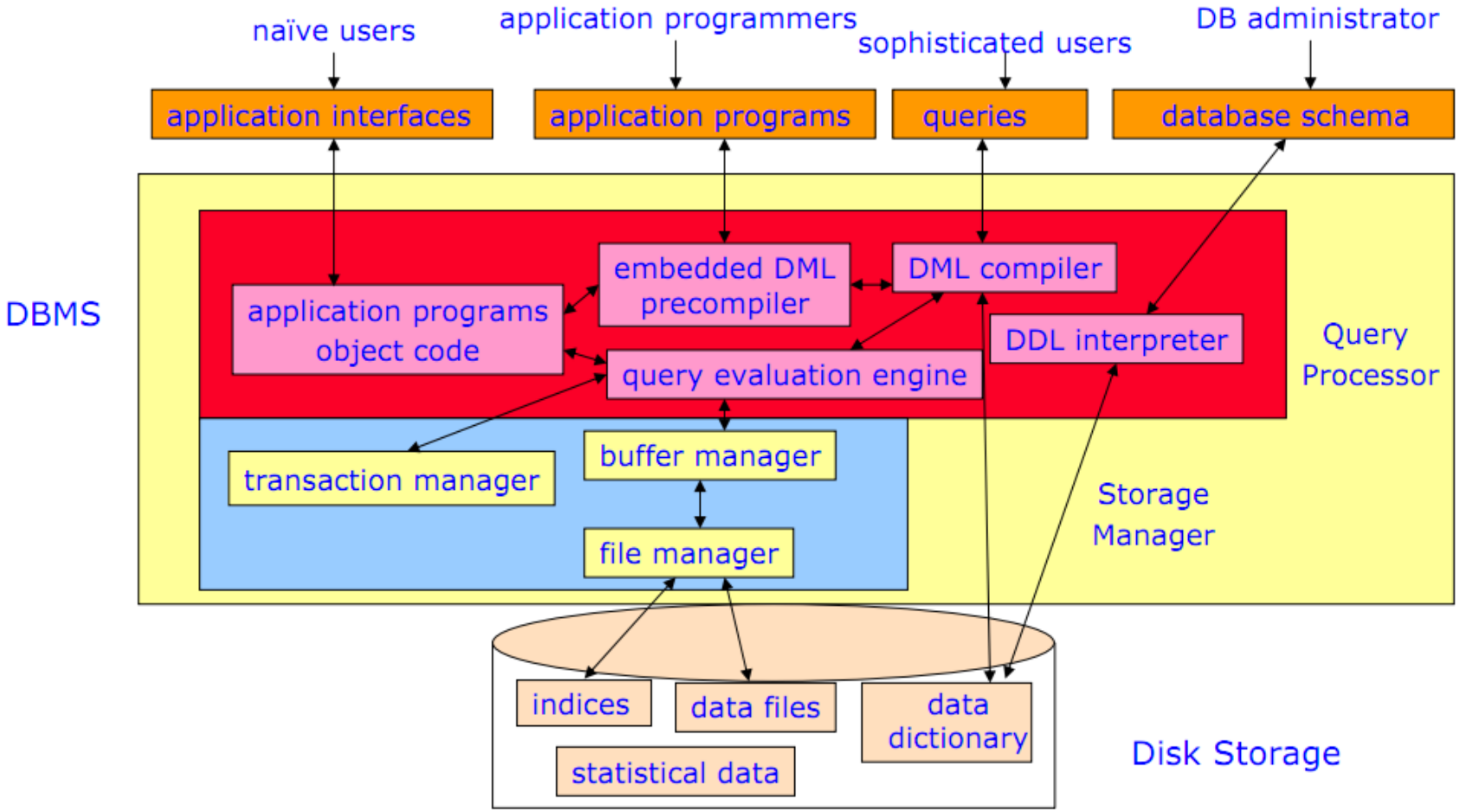
EECS3421 Topics

- The Relational Data Model
- Relational Algebra
- Defining, Querying and Manipulating Databases
  - the Structured Query Language (SQL)
- Application Programming with SQL
- Database Design and Normalization
- NoSQL Databases
- Advanced Topics (SQL security, Transactions, Recovery)

Why study databases?

- Shift from *computation* to *information*
  - always true for corporate computing
  - Web made this point for personal computing
  - more and more true for scientific computing
- Need for DBMS has exploded
  - Corporate: retail swipe/clickstreams, “customer relationship mgmt”, “supply chain mgmt”, “data warehouses”, etc.
  - Scientific: digital libraries, Human Genome project, Sloan Digital Sky Survey, physical sensors, grid physics network
- A practical discipline spanning much of
  - OS, languages, theory, AI, multimedia, logic
  - Yet with a focus on real-world apps

DBMS High-level Architecture



Advantages of a DBMS

- Data independence
- Efficient data access
- Data integrity & security
- Data administration
- Concurrent access, crash recovery
- Reduced application development time
- So why not use them always?
  - Expensive/complicated to set up & maintain
  - Cost & complexity must be offset by need
  - General-purpose, not suited for special-purpose tasks (e.g. text search!)

What comes next?

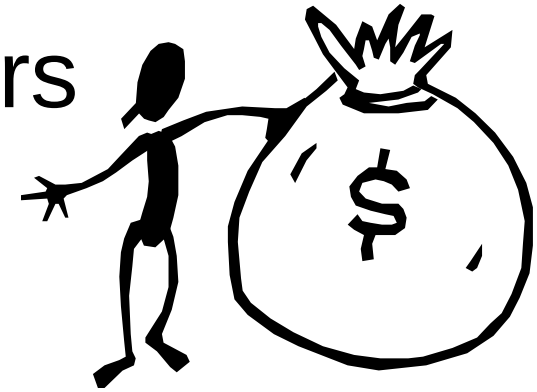
- If you are heading for industry:
  - Database professionals are in demand and well paid
- If you want to do research:
  - Many interesting problems ahead [The Beckman Report on Database Research, Oct 2013]
    - Scalable Big/Fast Data Infrastructures
    - Diversity in the Data Management Landscape
    - End-to-End Processing and Understanding of Data
    - Cloud Services
    - Roles of Humans in the Data Life Cycle
  - https://cacm.acm.org/magazines/2016/2/197411-the-beckman-report-on-database-research/fulltext
- Further studies in data systems at EECS:
  - eecs4411: Database Management Systems
  - eecs4415: Big Data Systems

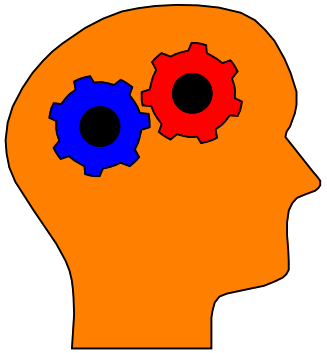
Summary (part 1)

- DBMS marries two very old concepts
  - The Database (idealistic vision)
  - File management system (imminently practical)
- Benefits
  - Maintain, query large datasets
  - Manipulate data and exploit semantics
  - Recover from system crashes
  - Juggle/balance concurrent access, automatic parallelization
  - Quick application development
  - Preserve data integrity and security
- Powerful abstractions provide data independence
  - Application safe from changes to data organization, hardware

Summary (cont.)

DB administrators, developers are the bedrock of the information economy





Data management R&D spans a broad, fundamental branch of the science of computation

This semester: become an effective DBMS user

COURSE ADMINISTRIVIA

Course administrivia

- Website:
  - https://www.eecs.yorku.ca/~papaggel/courses/eecs3421/
  - Read the course syllabus online!
- Discussion forum (Piazza):
  - https://piazza.com/yorku.ca/fall2019/eecs3421
  - Questions: to Piazza (so everybody benefits)
- Contact:
  - website and discussion board (Piazza) are required reading
  - personal matters: email me (include “eecs3421” in the subject)
- Moodle:
  - https://moodle.yorku.ca/moodle/course/view.php?id=165756
  - Used mostly for test assessment and grade distribution
- Office hours:
  - Mon, 13:00-14:00 online (same Zoom link)

Course prerequisites

- LE/EECS 2030 3.00 or LE/EECS 1030 3.00

Active Lectures

- Goal: get your gears turning in class
- Some in-class activities like:
  - problem solving, short quizzes
- A number of tutorials and online resources

Active Lectures (cont.)

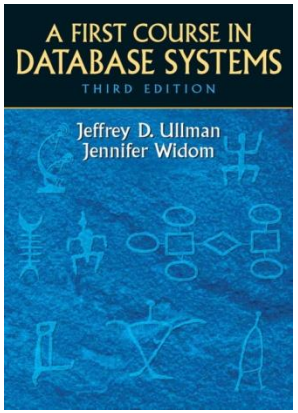
- Benefits of active learning
  - Exercise your knowledge and skills in class, with support
  - We'll know where the difficulties are
  - Get more from when I'm lecturing
- What it requires
  - Being active in class, including working with others and looking at each other's solutions to problems
  - A positive, encouraging environment

Course Marking Scheme

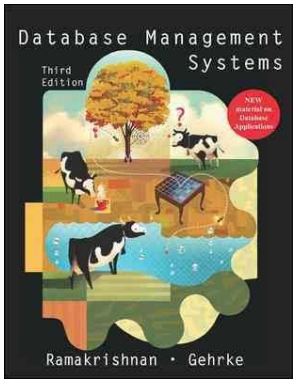
Work	Weight	Comment
3 Assignments	45%	15% each
1 Midterm Test	20%	-
Final Exam	35%	You must get >=40% to pass the course

Textbook

- A First Course in Database Systems, by Jeffrey D. Ullman and Jennifer Widom, 2008 (3rd Edition)



- (Optional) Database Management Systems, by Raghu Ramakrishnan and Johannes Gehrke, 2003 (3rd Edition)



Assignment Policies

- You may work with a partner on assignments
  - Can change partners between assignments
  - You may not dissolve a partnership in an assignment without permission
- Assignments must be submitted via *submit*
  - Your code must run on our lab computers (“PRISM”)
- Late policy
  - No late assignments will be accepted!
  - Submit on time!

Next Hour

- Relational Model