CMSC424: Database Design
Introduction/Overview

Today

- Administrivia

- Motivation:
  - Why study databases?
    - What are databases?

- Current Industry Outlook

- A typical DBMS at a glance
Logistics

- Professor: Peter Keleher
  - 5146 Iribe Bldg
  - keleher@umd.edu
  - Class Webpage:
    - http://sedna.cs.umd.edu/424

- Communication:
  - Piazza
  - Office hours. (starting Monday Sept 4)
  - Email to me: include 424 in subject as a last resort.
  - Do not message me on ELMS, I will not read it.

Logistics

All grades will be on grades.cs.umd.edu.

38% Weekly Assignments
We have 9 graded assignments:
- Assignments weights are 1,2: 3%, 3,4,5: 4%, 6,7,8,9: 5%.
- All are due Friday at midnight.

50% Exams
We have (3) exams:
- Test 1 is 15%.
- Test 2 is 15%.
- Final Exam is 20%.

12% Weekly Quizzes
12 weekly quizzes:
- Each is worth 1%.
- All are due Monday at midnight.
Logistics

Grading
- Whole class is curved: avg is B min, stdev up or down for A, C
- Approximate cut-offs last year (not guaranteed)
  - 85+: A-
  - 75+: B-
  - 65+: C-
  - 60+: D/F

Most had 40+ points (out of 50) on non-exams last year
- Exams are usually somewhat harder (no curves)
- Must average a passing grade on the total exam score

Web site: https://sedna.cs.umd.edu/424
Discussion: https://piazza.com/class/llhbqncu2r363c
Grades: https://grades.cs.umd.edu
Gradescope: https://www.gradescope.com/courses/535193
  - quizzes, assignment submissions, graded exams

Office Hours
- TAs (hours TBD)
  - projects
- Pete (me) IRB 5146, by appt
  - Everything else

ELMS
- Nope!
Some To-Dos

- Sign up for Piazza!
  - If not already added

- Set up the computing environment (Assign. 0), and make sure you can run Vagrant+VirtualBox, PostgreSQL, IPython, etc.

- Upcoming:
  - Quiz 1 (due Monday 9/4 midnight),
  - Assign 0: Environment. (this Friday, but not graded/no submit)
  - Assign 1: SQL. (next Friday, midnight)

Motivation: Data Overload

- Explosion of data, in pretty much every domain
  - Sensing devices and sensor networks that can monitor everything 24/7 from temperature to pollution to vital signs
  - Increasingly sophisticated smart phones
  - Internet, social networks makes it easy to publish data
  - Scientific experiments and simulations produce astronomical volumes of data
  - Internet of Things
    - Datafication: taking all aspects of life and turning them into data (e.g., what you like/enjoy turned into a stream of your "likes")

- How to handle that data? How to extract interesting actionable insights and scientific knowledge?

- Data volumes expected to get much worse
Four V’s of Big Data

- Increasing data Volumes
  - Scientific data: 1.5GB/genome -- can be sequenced in .5 hrs; LHC generates 100TB of data a day
  - 500M tweets per day (as of 2013)
  - As of 2012: 2.5 Exabytes of data created every day
  - EBay: Two data warehouses with 7.5PB and 40PB
  - Walmart: 583 terabytes of sales and inventory data
  - FICO monitors 2.5 billion active accounts worldwide

- Variety:
  - Structured data, spreadsheets, photos, videos, natural text, ...

- Velocity

- Veracity
  - Sensors everywhere -- can generate tremendous volumes of "data streams"
  - Real-time analytics requires data to be consumed as fast as it is generated

- Veracity
  - How do you decide what to trust? How to remove noise? How to fill in missing values?
Big Data and Data Science to the Rescue

- Terms increasingly used synonymously: also data analytics, data mining, business intelligence
  - Loosely used for any process where interesting things are inferred from data
  - Google search: “How Big Data Will Change”

- Data scientist called the sexiest job of the 21st century
  - The term has becoming very muddled at this point

- Overhyped words
  - We are headed toward the trough of Disillusionment

Is it all hype?

- No: Extracting insights and knowledge from data very important, and will continue to increase in importance
  - Big data techniques are revolutionizing things in many domains like Education, Food Supply, Disease Epidemics, ...

- But: it is not much different from what we, especially statisticians, have been doing for many years

- What is different?
  - Much more data is digitally available than was before
  - Inexpensive computing + Cloud + Easy-to-use programming frameworks = Much easier to analyze it
  - Often: large-scale data + simple algorithms > small data + complex algorithms
    - Changes how you do analysis dramatically
Motivation: Data Overload

- How do we do anything with this data?

- Where and how do we store it?
  - Disks are doubling every 18 months or so -- not enough
  - In many cases, the data is not actually recorded as it is; *summarized* first

- What if the disks crash?
  - Very common, especially with 10,000’s of disks

- How do we ensure “correctness”?
  - What if the system crashes in the middle of an ATM transaction?
    - Can’t have money disappearing
  - What happens when a million people try to buy tickets to *your favorite artist’s concert* at the same time?

Motivation: Data Overload

- What to do with the data? How to process/analyze it?
  - text search?
    - Very limited
  - “find the stores with the maximum increase in sales in last month”
    - We can’t expect the users to write Java programs
  - “how much time from here to Pittsburgh if I start at 2pm?”
    - Data is there; more will be soon (GPS, live traffic data)
    - Requires predictive capabilities
  - Increasing need to convert “information” to “knowledge”: Data mining
    - “How should we replicate different movies?” (Netflix)
    - Find videos with this type of an event (say car break-ins)
    - Mine the “blogs” to detect “buzz”
Motivation: Data Overload

- Speed !!
  - With TB’s of data, just finding something (even if you know what), is not easy
    - Reading a file with TB of data can take hours
  - Imagine a bank and millions of ATMs
    - How much time does it take you to do a withdrawal ?
    - The data is not local

- How do we guarantee the data will be there 10 years from now ?

- Privacy and security !!!
  - Every other day we see some database leaked on the web
    - identity fraud, influencing elections...
  - How to make sure different users’ data is protected from each other

Why not use file systems ?

- Drawbacks of using file systems to store data:
  - Data redundancy and inconsistency
    - Multiple file formats, duplication of information in different files
  - Difficulty in accessing data
    - Need to write a new program to carry out each new task
  - Data isolation — multiple files and formats
  - Integrity problems
    - Integrity constraints (e.g., account balance > 0) become “buried” in program code rather than being stated explicitly
    - Hard to add new constraints or change existing ones
Why not use file systems?

- Drawbacks of using file systems to store data:
  - Atomicity of updates
    - Failures may leave database in an inconsistent state with partial updates carried out
    - Example: Transfer of funds from one account to another should either complete or not happen at all
  - Concurrent access by multiple users
    - Concurrent access needed for performance
    - Uncontrolled concurrent accesses can lead to inconsistencies
    - Example: Two people reading a balance (say 100) and updating it by withdrawing money (say 50 each) at the same time
  - Security problems
    - Hard to provide user access to some, but not all, data

DBMSs to the Rescue

- Provide a systematic way to answer many of these questions...
- Aim is to allow easy management of high volumes of data
  - Storing, Updating, Querying, Analyzing....

- What is a Database?
  - A large, integrated collection of (mostly structured) data
  - Typically models and captures information about a real-world enterprise
    - Entities (e.g. courses, students)
    - Relationships (e.g. John is taking CMSC 424)
  - Usually also contains:
    - Knowledge of constraints on the data (e.g. course capacities)
    - Business logic (e.g. pre-requisite rules)
    - Encoded as part of the data model (preferable) or through external programs
DBMSs to the Rescue

Massively successful for *highly structured data*

- Why? Structure in the data (if any) can be exploited for ease of use and efficiency
  - If there is no structure in the data, hard to do much
  - Contrast managing emails vs managing photos
- Much of the data we need to deal with is highly structured
- Some data is *semi-structured*
  - E.g.: Resumes, Webpages, Blogs etc.
- Some has complicated structure
  - E.g.: Social networks
- Some has no structure
  - E.g.: Text data, Video/Image data etc.