

Queries for Today

Database Management Systems

Fall 2016

“Knowledge is of two kinds: we know a subject ourselves, or we know where we can find information upon it.”

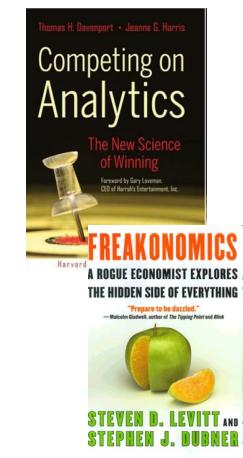
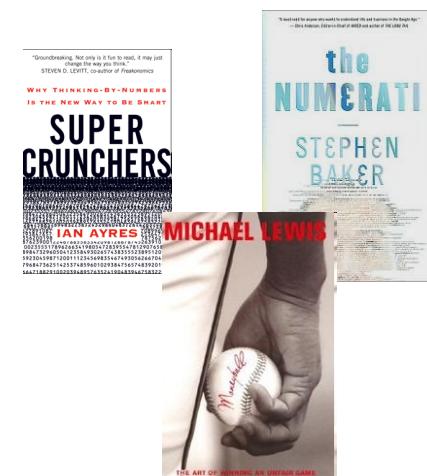
-- Samuel Johnson (1709-1784)

- Why?
- Who?
- What?
- How?
- For instance?

Databases – Why Study Them?



Databases – Why Study Them?



The “Big Data” Buzz – Why?

“Between the dawn of civilization and 2003, we only created five exabytes of information; now we’re creating that amount every two days.” Eric Schmidt, Google (and others)



Search Images Mail Documents Calendar Sites Groups More

Google

you tube cat videos

Search About 124,000,000 results (0.15 seconds)

Web

Images

Maps

Videos

News

Shopping

More

Oakland, CA

Change location

Any duration

Short (0 - 4 min.)

Medium (4 - 20 min.)

Long (20+ min.)

More search tools

Probably the Funniest Cat Video You'll Ever See



www.youtube.com/watch?v=SU...
Jan 12, 2007 - 3 min - Uploaded by...
Now don't let the corny opening fo...
hilarious **cat video** you will ever see

Supercats: Episode 1 — The Funniest Cat Vid



www.youtube.com/watch?v=mf...
Jul 22, 2009 - 3 min - Uploaded by...
Download Cat Piano from iTunes: h...
Human-to-Cat Translator: http://bit.

The two talking cats - YouTube



www.youtube.com/watch?v=z3Uo...
Jun 28, 2007 - 55 sec - Uploaded by...
Alert icon: You need Adobe Flash P...
Standard YouTube License ... Self

10 Cutest Cat Moments - YouTube



www.youtube.com/watch?v=q1dp...
Mar 6, 2009 - 6 min - Uploaded by...
The clips for this compilation of cut...
our favorite **videos** ... Standard Ya

More videos for you tube cat videos »

Top 10 Funny Cat Videos on YouTube

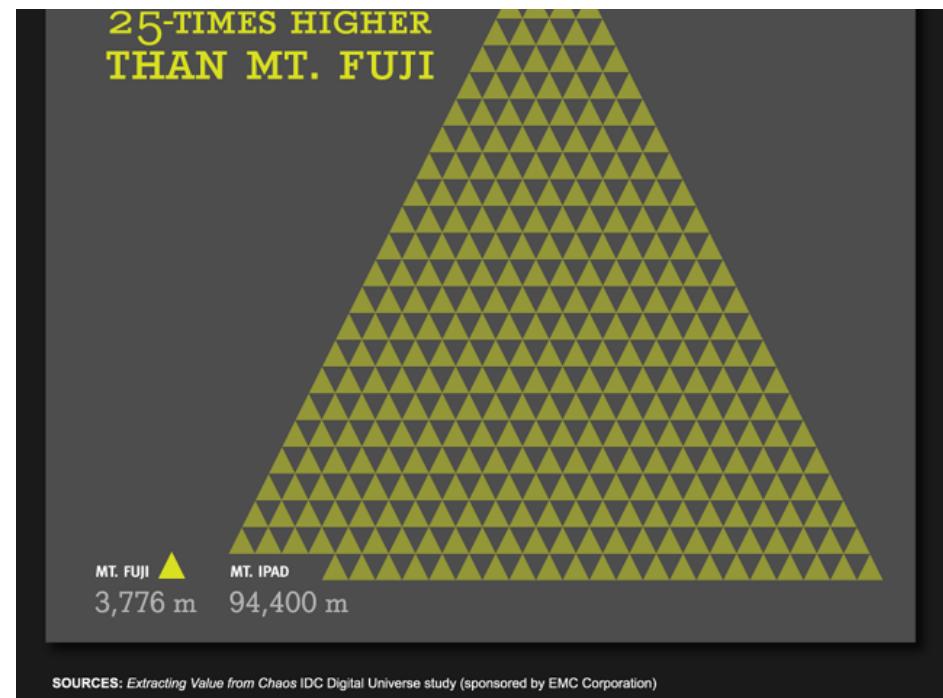
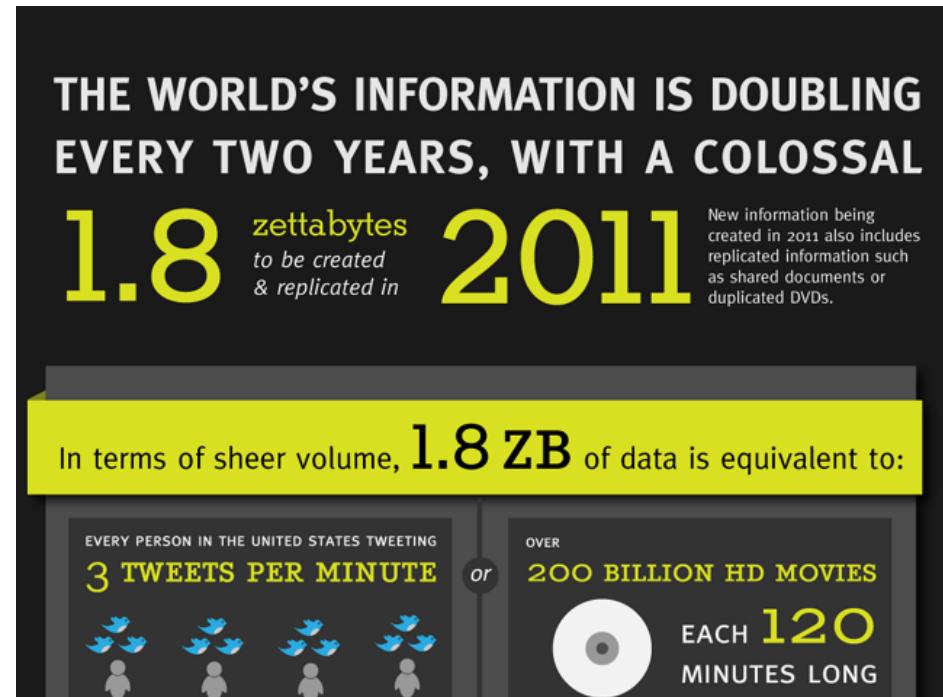


mashable.com/2010/04/17/funny-cat-videos-youtube/
by Amy Mae Elliott - in 16,907 Google+ circles
Apr 7, 2010 - We've already brought you ten hilarious **cat** clips, but dogs shouldn't be the only ones to have

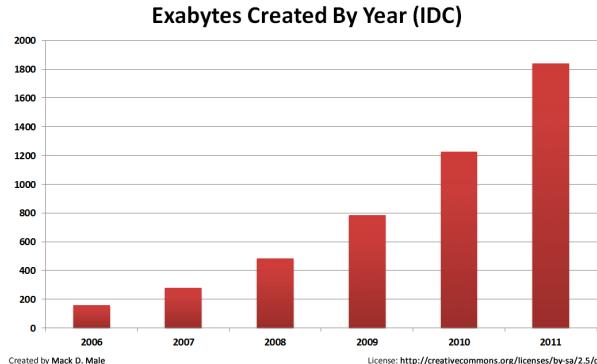
Storing **1.8 ZB** of information would take:

**57.5 BILLION
32GB APPLE IPADS**

WITH THAT MANY iPADS WE COULD
BUILD A MOUNTAIN OF iPADS THAT IS
**25-TIMES HIGHER
THAN MT. FUJI**



The “Big Data” Buzz – Why?



SI decimal prefixes		Binary usage
Name (Symbol)	Value	
kilobyte (kB)	10^3	2^{10}
megabyte (MB)	10^6	2^{20}
gigabyte (GB)	10^9	2^{30}
terabyte (TB)	10^{12}	2^{40}
petabyte (PB)	10^{15}	2^{50}
exabyte (EB)	10^{18}	2^{60}
zettabyte (ZB)	10^{21}	2^{70}
yottabyte (YB)	10^{24}	2^{80}

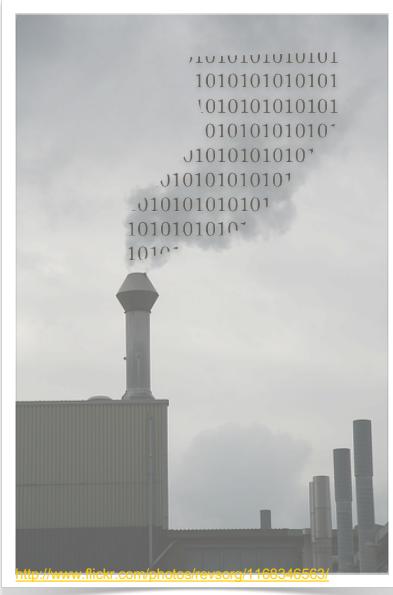
“The sexy job in the next 10 years will be statistician.” “Data Scientists”



Hal Varian
Prof. Emeritus UC Berkeley
Chief Economist, Google

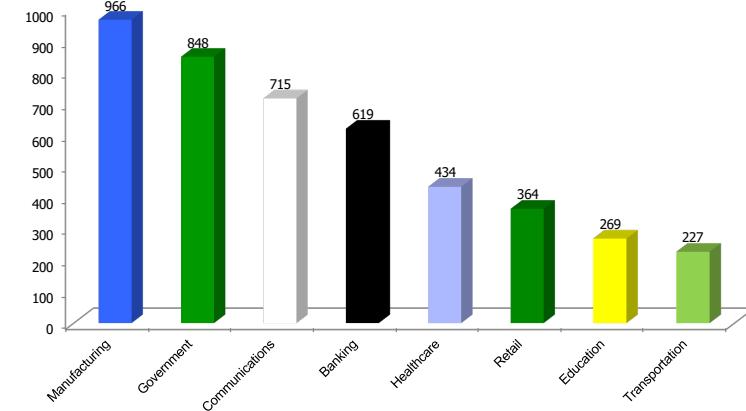
Industrial Revolution of Data!

- **UPC**
- **RFID**
- **GPS**
- **Sensornets**
- **Software Logs**
- **Microphones**
- **Cameras**
- ...



Some Numbers by Industry

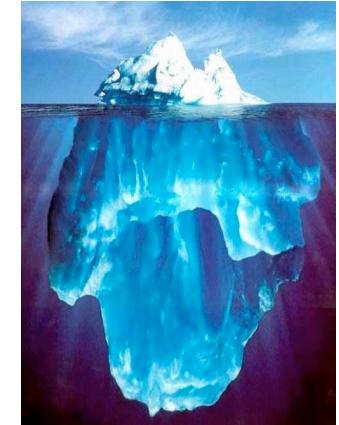
Amount of Stored Data By Sector
(in Petabytes, 2009)



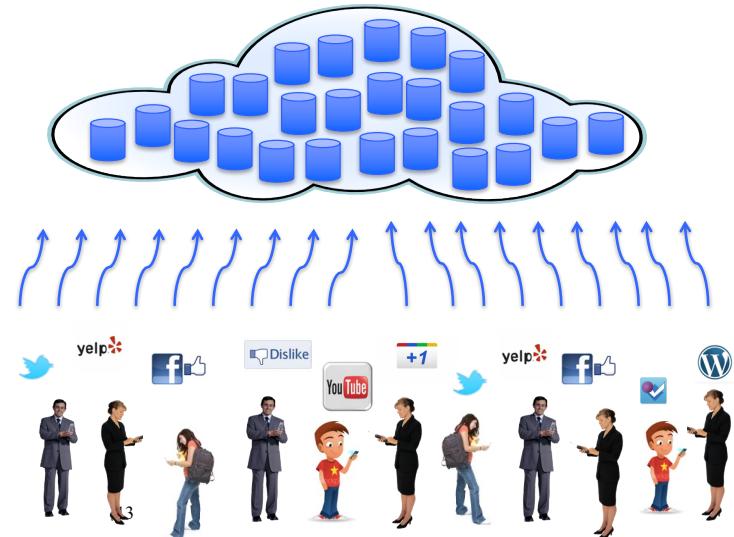
Sources:
"Big Data: The Next Frontier for Innovation, Competition and Productivity."
US Bureau of Labor Statistics | McKinsey Global Institute Analysis

It's All Happening On-line

- **Every:**
 - Click
 - Ad impression
 - Wall post, friending, ...
 - Billing event
 - Fast Forward, pause,...
 - Server request
 - Transaction
 - Network message
 - Fault
 - ...
- **Generates Streams of Data that can be Analyzed**

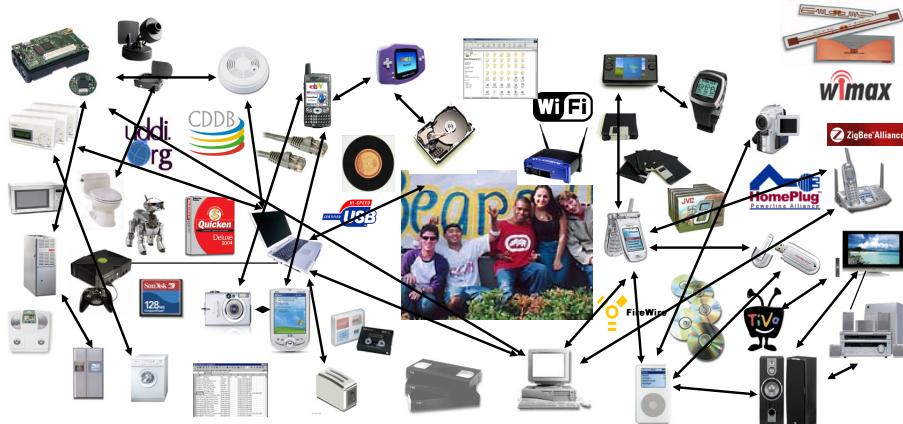


User Generated Content



Credit: Mike Carey, UCI

M2M - Internet of things



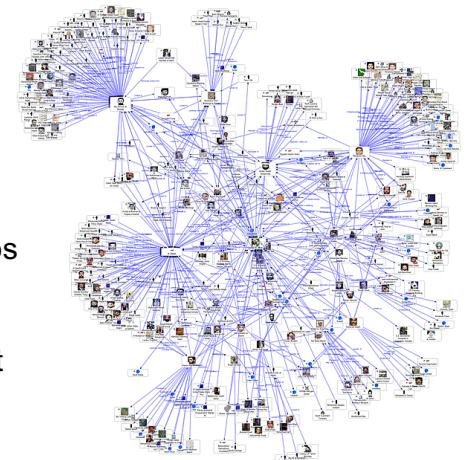
15

Graph Data

Lots of interesting data has a graph structure:

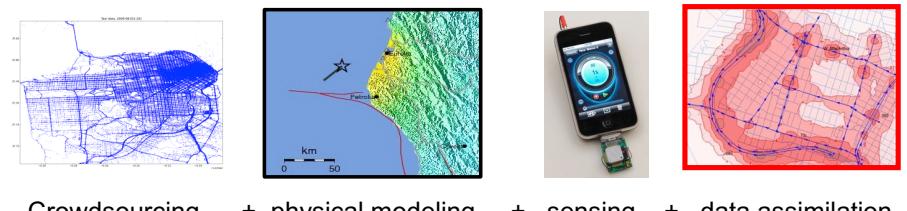
- Social networks
- Communication networks
- Computer Networks
- Road networks
- Citations
- Collaborations/Relationships
- ...

Some of these graphs can get quite large (e.g., Facebook's user graph)



14

Fusion: e.g., NextGen Maps



to produce:



16

What can you do with the data

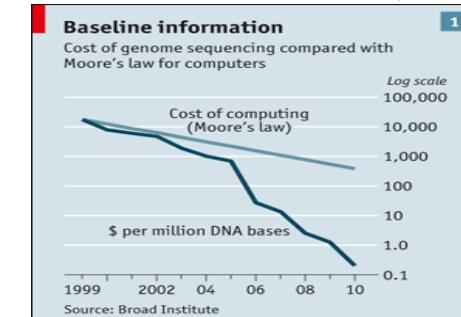
- **Reporting**
 - Post Hoc
 - Real time
- **Monitoring (fine-grained)**
- **Exploration**
- **Finding Patterns**
- **Root Cause Analysis**
- **Closed-loop Control**
- **Model construction**
- **Prediction**
- ...

17

Big Data, Societal-Scale App?

- **Cancer Tumor Genomics**
- **Vision: Personalized Therapy**
 - "...10 years from now, each cancer patient is going to want to get a genomic analysis of their cancer and will expect customized therapy based on that information."

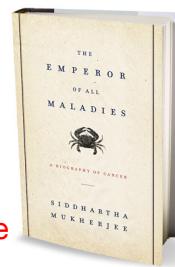
Director, The Cancer Genome Atlas (TCGA), *Time Magazine*,
6/13/11



Opportunity or Obligation?

- **Provocative Hypothesis:** Given fast growing genomic databases, could CS now be a huge help in war on cancer?
- If a *chance* that we could help millions of cancer patients live longer and better lives, as moral people, *aren't we obligated to try?*

David Patterson, "Computer Scientists May Have What It Takes to Help Cure Cancer," *New York Times*, 12/5/2011



- UCSF cancer researchers + UCSC cancer genetic database + AMP Lab

The Cancer Genome Atlas: 5 PB = 20 cancers x 1000 genomes

So, In Summary...Why?

Data will be at the center of the major issues and events of your life.

As a computer professional, you'd better be on top of how to manage, use, and make sense of it.

Who?

- **Haidar M. Harmanani**
 - Professor of Computer Science
 - Office: 810 Block A
 - Office Hours: TTh: 8:00-9:30 and 3:00-4:30
 - Email: haidar@lau.edu.lb
- **Course Website:**
 - <http://vlsi.byblos.lau.edu.lb/classes/csc375/csc375.html>

Queries for Today

- Why?
- Who?
- **What?**
- How?
- For instance?

What is a Database?

[The Relational Model] provides a basis for a high level data language which will yield maximal independence between programs on the one hand and machine representation on the other. (E.F. Codd, 1981 Turing Award winner)



In Other Words...

Relational DataBase Management Systems were invented to let you use one set of data in multiple ways, including **ways that are unforeseen** at the time the database is built and the 1st applications are written.

(Curt Monash, analyst/blogger)

That is, think about the data independently of any particular program.

What is a Database?

- Let's not split hairs.

A database is a large collection of structured data

What is a DBMS?

A Database Management System (DBMS) is software that stores, manages and/or facilitates access to databases.

- Traditionally, term used narrowly
 - Relational databases with transactions
- Warning: market and terms in rapid transition
 - The tech remains (roughly) the same
 - Good time to focus on fundamentals!

Some Ways you May Use Databases



Find the Database(s) in This Picture

The Pandora website interface includes:

- The header: PANDORA® internet radio, account, sign out, upgrade.
- The sidebar: Create a New Station..., Your Stations, Weather Report Radio, Pat Metheny Radio, add variety, options, Bon Iver Radio, Les Nubians Radio, Peter Gabriel Radio, Little Feat Radio, QuickMix.
- The main content area:
 - She by Christian Scott on: Rewind That
 - Like Minds by Gary Burton ... on: Like Minds
 - Episode D'azur by: Pat Metheny on: We Live Here
- Information about Pat Metheny: One of the most original guitarists from the '80s onward (he is instantly recognizable), Pat Metheny is a classical-trained player who has won popularity but also taken some wild left turns. His records with the Pat Metheny Group are difficult to describe (bebop-jazz? modern music?) but managed to be both accessible and original, stretching the boundaries of jazz and making Metheny famous.
- Call-to-action buttons: Buy, Bookmark, Share.
- Footer links: About This Music, Artist, Album, Song, Lyrics, Fans.
- Bottom navigation: Your Concert Listings, Find Shows Now, Your Local Shows, Station Gifting, Make a station for a friend, On Your Mobile Phone, Find Your Phone, Your Bookmarked Songs, Check Out.
- Right sidebar: Similar Artists, LOCAL CONCERTS by YOUR PANDORA ARTISTS, FIND SHOWS NEAR YOU, VISA SIGNATURE.

Other Kinds of Databases (spatial)

The screenshot shows the Yahoo! Local Maps interface. At the top, it says "Loading "Yahoo! Maps, Driving Directions, and Traffic"". Below the search bar, there's a "Sign In" link and a "Search" button. On the left, there's a sidebar titled "FIND A BUSINESS ON THE MAP" with a search input for "uc berkeley" and a "Search" button. The main area is a map of Berkeley, California, with numerous streets labeled. Twenty-four numbered orange circles (1-24) are placed at specific locations across the city, likely marking points of interest or businesses.

Other Kinds of Databases (genome)

The screenshot shows the NCBI Map Viewer interface for the Homo sapiens genome. At the top, it says "on chromosome(s)" and has a "Find" button. The main area is titled "Homo sapiens genome view" and "Build 35.1 statistics". It displays a series of vertical bars representing chromosomes, numbered 1 through 22, X, and Y. The bars are black with white internal segments, indicating gene density or other genomic data. To the right, there's a "BLAST search the human genome" link. The left sidebar includes links for "Map Viewer Home", "Related Resources", and "Sequence Data".

Databases Make Life Better?

- “Players could finally sign up for the Star Wars Galaxies game last week as Sony opened up registration to the public.”
- “Once players got in to the game they found that the game servers were offline because of **database problems**.”
- “Some players spent hours tuning their in-game characters only to find that **crashes deleted all their hard work**.”



Source: BBC News Online, July 1, 2003.

What: Is an OS a DBMS?

- **Data can be stored in RAM**
 - every programming language offers this
 - RAM is fast, and random access
 - isn't this heaven?

What: Is an OS a DBMS?

- **Data can be stored in RAM**
 - every programming language offers this
 - RAM is fast, and random access
 - isn't this heaven?
- **Every OS includes a File System**
 - manages *files* on a magnetic disk
 - allows *open, read, seek, close* on a file
 - allows protections to be set on a file
 - drawbacks relative to RAM?

What: File System vs DBMS

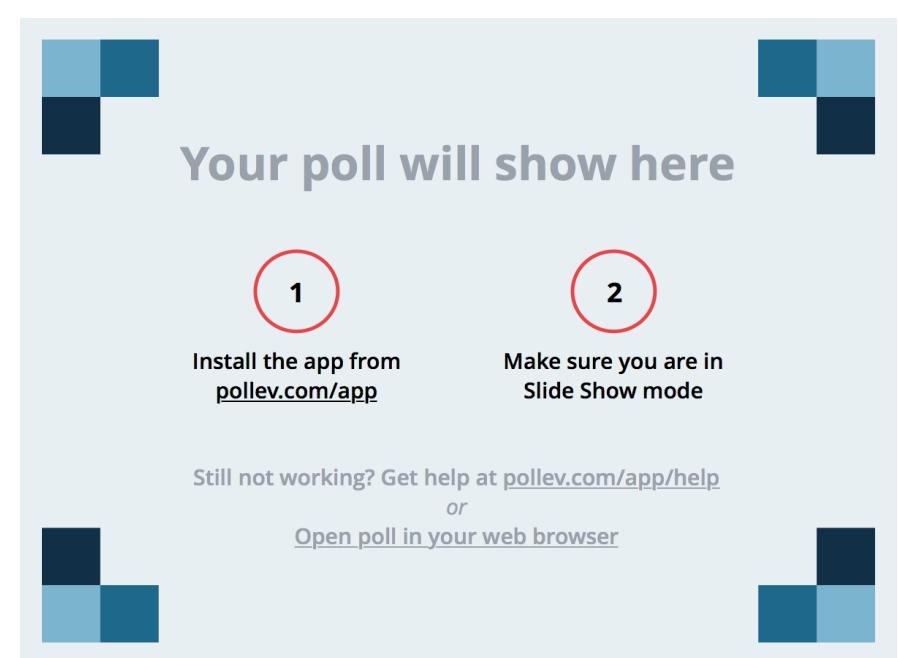
• Thought Experiment 1:

- You're updating a file.
- The power goes out.
- Which changes survive?

- a) all
- b) none
- c) all since last save
- d) ???

• Answer to: PollEv.com/harmanani765

What: File System vs DBMS

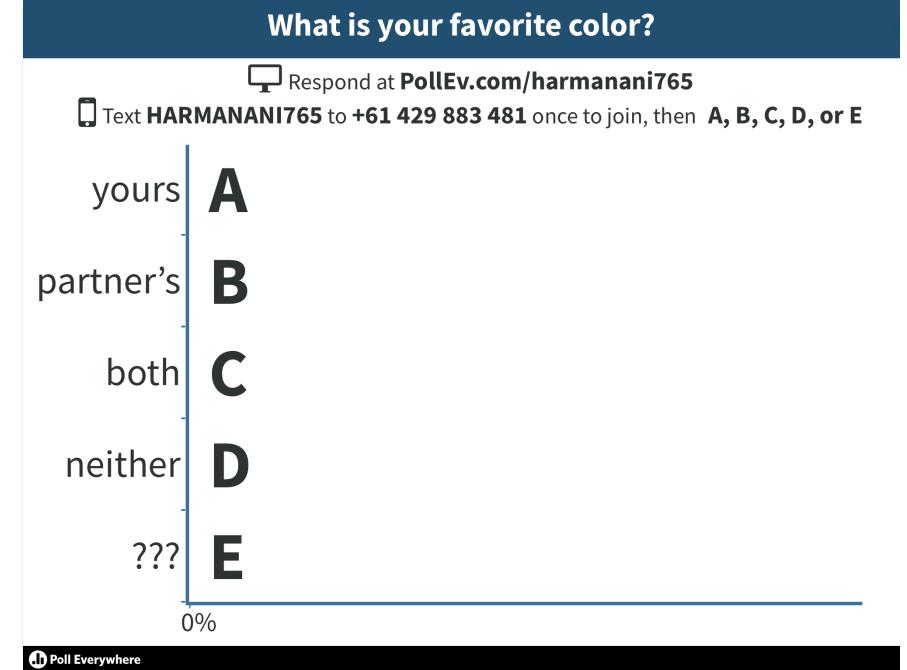


What: File System vs DBMS

- **Thought Experiment 2:**

- You and your project partner edit the same file.
- You both save it at the same time.
- Whose changes survive?

- a) yours
- b) partner's
- c) both
- d) neither
- e) ???



What: File System vs DBMS

- **Thought Experiment 1:**

- You and your project partner edit the same file

Q: How do you code against an API that guarantees you “???” ?

A: *Very carefully.*

- The power goes out.
- Which changes survive?

- a) all
- b) none
- c) all since last save
- d) ???

What: Database Systems

- **What more could we want than a file system?**

- Clear *API contracts* regarding data
 - concurrency control, replication, recovery
- Simple, efficient, well-defined *ad hoc*¹ queries
- Efficient, scalable bulk processing
- Benefits of good data modeling

- **S.M.O.P.²? Not really...**

¹ad hoc: formed or used for specific or immediate problems or needs

²SMOP: Small Matter Of Programming

What: Current Market

- Relational DBMSs anchor the software industry
 - Elephants: Oracle, IBM, Microsoft, Teradata, HP, EMC, ...
 - Open source: MySQL, PostgreSQL
- Obviously, Search
 - Google & Bing
- Open Source “NoSQL”
 - Hadoop MapReduce
 - Key-value stores: Cassandra, Riak, Voldemort, Mongo, ...
- Cloud services
 - Amazon, Google AppEngine, MS Azure, Heroku, ...
- Increasing use of custom code



So... What Is a Database?

- **Database:**
An (often) large, integrated collection of data.
- **Models a real-world *enterprise***
 - Entities (e.g., teams, games)
 - Relationships
(e.g., Cal *plays against* Stanford *in* The Big Game)
 - Can also include active components , often called “business logic”. (e.g., the BCS ranking system)

Key Concept: Structured Data

- A data model is a collection of concepts for describing data.
- A schema is a description of a particular collection of data, using a given data model.
- The relational model of data is the most widely used model today.
 - Main concept: relation, basically a table with rows and columns.
 - Every relation has a schema, which describes the columns, or fields.

Example: University Database

- **Conceptual schema:**
Students(sid: string, name: string, age: integer, gpa:real)
Courses(cid: string, cname:string, credits:integer)
Enrolled(sid:string, cid:string, grade:string)
 FOREIGN KEY sid REFERENCES Students
 FOREIGN KEY cid REFERENCES Courses
- **External Schema (View):**
Course_info(cid:string,enrollment:integer)
Create View Course_info AS
SELECT cid, Count (*) as enrollment
FROM Enrolled
GROUP BY cid

e.g.: An **Instance** of Students Relation

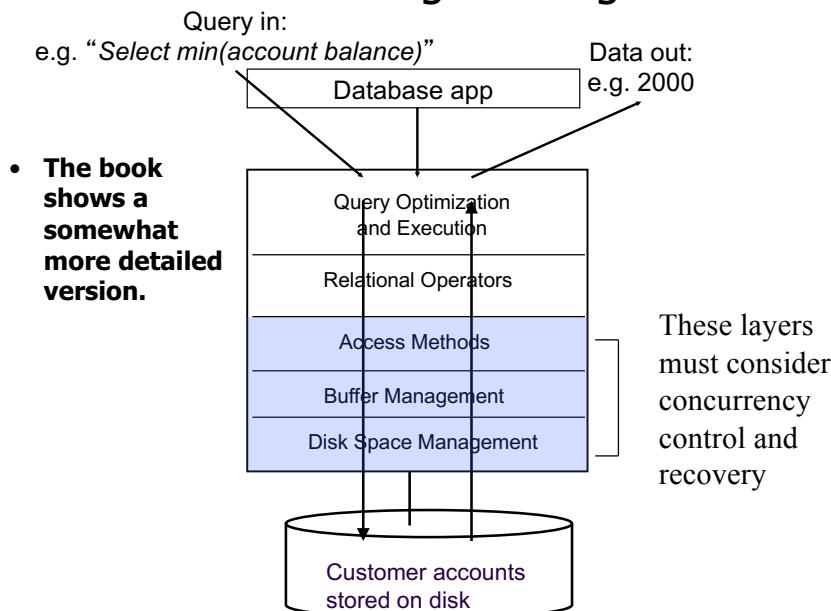
sid	name	login	age	gpa
53666	Jones	jones@cs	18	3.4
53688	Smith	smith@eecs	18	3.2
53650	Smith	smith@math	19	3.8

What is a Database System?



- A **Database Management System (DBMS)** is a software system designed to **store, manage, and facilitate access to databases**.
- **A DBMS provides:**
 - Data Definition Language (DDL)
 - Data Manipulation Language (DML)
 - Queries – to retrieve, analyze and modify data.
 - Sometimes called “CRUD”
 - Guarantees about durability, concurrency, semantics, etc.
- **Three main uses: Transactional, Archival, and Analytical**

A DBMS “Lasagna” Diagram



What: Current Market

- **Relational DBMSs anchor the software industry**
 - Elephants: Oracle, IBM, Microsoft, Teradata, HP, EMC, ...
 - Open source: MySQL, PostgreSQL
- **Obviously, Search**
 - Google & Bing
- **Open Source “NoSQL”**
 - Hadoop MapReduce
 - Key-value stores: Cassandra, Riak, Voldemort, Mongo, ...
- **Cloud services**
 - Amazon, Google AppEngine, MS Azure, Heroku, ...
- **Increasing use of custom code**

What will we learn?

- Design patterns for dealing with Big Data
- When, why and how to structure your data
- How MySQL and Oracle and (a bit of) Google work
- SQL ... and noSQL
- Managing concurrency
- Fault tolerance and Recovery
- Scaling out: parallelism and replication
- Audacity and Reverence.

What: Summing up

- Data is at the center of many things.
- For instance: computer science.

What: Summing up

You might think that we'll learn to apply computer science to Big Data.

The techniques we'll learn for Big Data are the key to scalable computer science.

Don't forget Hal Varian's prediction...

Google's Chief Economist

- **These professions barely have names:**
 - Cloud programmer
 - Data scientist
 - Scalable systems architect
 - Data-driven thinker
- **In 5 years, this will be a large fraction of the computing workforce.**

“ By 2018, the US could face a shortage of up to 190,000 workers with analytical skills” McKinsey Global Institute