CMPS 180: Database Systems I

Fall 2017

Instructor: Sheldon (Shel) Finkelstein

Textbook:

"A First Course in Database Systems", Jeffrey Ullman and Jennifer Widom, Prentice-Hall, 3rd edition.

Science and Engineering Library has textbook on Reserve.

2cd edition is also available, but this course uses 3rd edition.

Course Syllabus and Chapter 1

CMPS 180, Fall 2017 Course Information

Classes: Mon, Wed, Fri, 2:40 - 3:45PM, BE-152
Instructor: Sheldon (Shel) Finkelstein, shel@ucsc.edu
Office Hours: Friday 9:00 - 10:00am, or by appointment, E2-345B
(Note that I changed offices this term.)

Teaching Assistants:

Akhil Dixit, akadixit@ucsc.edu
Aniket Kulkarni, amkulkar@ucsc.edu

Labs Sections (all in BE-105)

01A: Tues 9:00 - 10:40AM (Akhil)
01B: Wed 10:00 - 11:40AM (Akhil)
01C: Thu 6:00 - 7:40PM (Aniket)

Piazza Signup: https://piazza.com/ucsc/fall2017/cmps180
Course Page (including slides after lectures):
https://piazza.com/ucsc/fall2017/cmps180/home

Some Background: Shel Finkelstein

- IBM Almaden Research: Database and Distributed Systems
- Tandem Computers: Transaction Management and System Managed Storage
- Illustra Information Systems: Productizing Postgres (Time Series, Administration)
- ADB/Matisse (Object Database): Chief Scientist
- Sun Microsystems: Managed Enterprise Java architecture and partnership relationships when Java Enterprise Edition was created
- SAP: VP, Research Fellow, Chief Tech Architect, focusing on Applications and Database
- Teaching at UC Santa Cruz since January 2014, mainly Database courses

Teaching Assistants

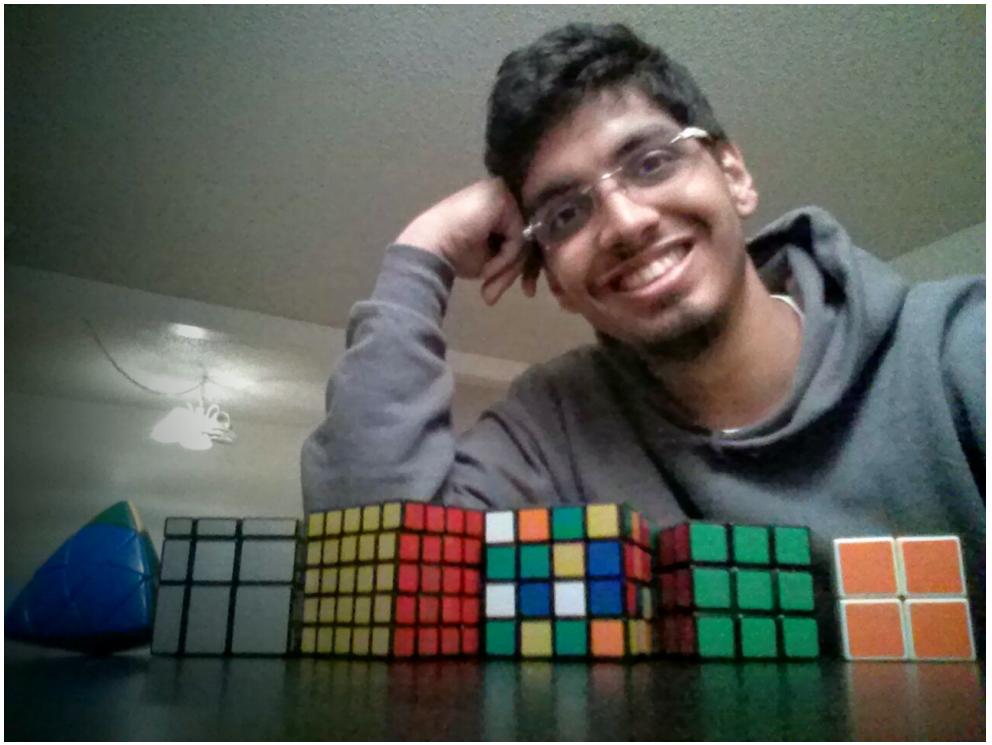
Akhil Dixit
akadixit@ucsc.edu

Lab Sections (in BE-105)

01A: Tues 9:00 - 10:40AM
01B: Wed 10:00 - 11:40AM

Office Hours:

Tues 11:00am-noon, **BE-119**
(or by appointment)



Teaching Assistants

Aniket Kulkarni
amkulkar@ucsc.edu

Lab Section (in BE-105)

01C: Thu 6:00 - 7:40PM

Office Hours:

Wed 1:30pm to 2:30pm, **BE119**
(or by appointment)

Course Description/Syllabus

CMPS 180 covers concepts, approaches, tools, and methodology of database design and querying. Topics include:

- Relational Data Model and its history
- SQL language: Data Definition, Queries and Updates, Indexes, Views, Constraints, Rules
- Relational Algebra, Query Execution and Transaction Processing
- Database Application Development
- Design Theory: Schemas and Normal Forms
- On-Line Analytical Processing (OLAP)
- Semi-Structured Data Models: XML, JSON, etc.
- Cloud Databases (time permitting)
- Big Data and NOSQL (time permitting)

Lots of practical material and logic … and lots of concepts and theory, particularly in the second half of the term

Tentative Lecture Schedule

Topic	# Lectures	Dates	Chapters
History and Introduction	1	9/29	1
The Relational Data Model	2	10/2 – 10/4	2.1, 2.2
SQL: DDL, DML; defining relations & constraints, and writing queries	10	10/6 - 10/27	2.3, 2.5, 6.1-6.5, 7.1, 7.2
Relational Algebra and Query Execution	3	10/30 – 11/3	2.4, 2.5
Midterm	1	Monday 11/6	-
Database Application Development	2	11/8 – 11/13	9.1, 9.2, 9.6
Schema Refinement and Normal Forms	5	11/15 – 11/27	3.1-3.5
OLAP and OuterJoin	1	11/29	10.6, 5.2.7
Semi-Structured Data Models: XML, JSON, etc.	2	12/1 – 12/4	11.1-11.3, 12.1-12.2
Big Data and NOSQL	1	12/6	
Review (time permitting)	1	12/8	
	`		
Final, in classroom		Wed 12/13, noon – 3pm	

Course Evaluation

- Gradiance Homeworks 10%
- Lab Assignments 20%
- Midterm 30%
- Final Exam 40%

General Information for CMPS 180 appears in Piazza on Course Information page; see also the General Information file that's under

Resources → Resources→General Resources

Lectures

- You are responsible for all material presented in Lectures.
  - Some material might not be in the textbook.
  - Some information might not be on the slides.
- Lecture slides will be posted on Piazza under Resources-->Lectures
  - Slides may be modified after initial posting.
  - Date shown on Piazza is date that the Lecture presentation began.
- You should learn the concepts, not just learn what's presented in the Lectures.

Homeworks: Labs

- This course involves database application development projects through 4 Lab Assignments.
  - You must attend a Lab Section, where the Teaching Assistant will help guide you through Lab Assignments and answer your questions.
  - Lab Assignments will be submitted via Canvas, as zip files.
  - You'll have to learn how to move files from unix to your computers, so that you can post them on Canvas.
- We'll also use Canvas for grading.
  - Login to Canvas at https://canvas.ucsc.edu using your CruzID and Gold password. CMPS 180 should be one of the classes available.
  - Info on Canvas is at http://its.ucsc.edu/canvas/index.html

Tentative Lab Assignment Dates

- **Lab Sections** are a required part of course.
  - There **will** be Labs during the first week of classes, introducing you to Gradiance and PostgreSQL.
- **Lab1**
  - Assigned Monday, October 2
  - Due Sunday October 15
- **Lab2**
  - Assigned Monday October 16
  - Due Sunday October 29
- **Lab3**
  - Assigned Monday October 30
  - Due Sunday November 19 (extra week--Midterm is November 6)
- **Lab4**
  - Assigned Monday, November 20
  - Due Sunday December 3
- **Last day of classes**: Friday, December 8
  - During last week of classes, Lab Sections will be held, discussing answers to Lab4, as well as addressing questions about overall course before Final Exam.

Other Homeworks

- Non-lab homework assignments will be assigned via Gradiance, which has automated grading; see: http://www.gradiance.com/pub/stud-guide.html
  - For CMPS180, go to http://www.gradiance.com/services to enroll in this class, using class code **DF53BA19**
  - Gradiance uses HTTP, rather than HTTPS, so be sure to use a password that you don't use for any other account!
- There will also be ungraded Practice Homeworks that you don't hand in.
  - You're responsible for those Practice Homeworks, even though we won't discuss most of them in class.

Exams

- Midterm on Monday, November 6 (65 minutes long) covers first half of term, including material through Friday, November 3.

- Final Exam on Wednesday, December 13, noon – 3pm (3 hours) is comprehensive, covering material from entire term, with greater emphasis on second half of term.

- For both the Midterm and the Final Exam, you may bring in an 8.5 by 11 sheet of paper, with anything that you can read unassisted printed or written on both sides of the paper.
  - Sheets may not be shared during the exams.
  - You must show UCSC ID when you hand in your exams.

- **No** Make-up Exams. **No** Early/Late Exams.

Learning Support Center and Modified Supplemental Instruction

- UCSC Learning Support Services: http://lss.ucsc.edu/
  - There will be a CMPS 180 Learning Assistant this term, Alexander Ou, who will introduce himself in class.
  - Students may sign-up for tutoring at the Slug Success website: https://sserc.ucsc.edu/slug-success beginning Friday, October 6th at 5:00PM
  - Slots tend to fill up, and are usually not available if you try to sign-up before Exams or due dates. If you're interested in getting tutoring, please consider signing up early.

Interim Title IX Disclosure Statement

- Please be aware that under the UC Policy on Sexual Violence and Sexual Harassment, faculty and student employees (including Teaching Assistants, Readers, Tutors, etc.) are "responsible employees" and are **required** to notify the Title IX Officer of any reports of incidents of sexual harassment and sexual violence (sexual assault, domestic and dating violence, stalking, etc.) involving students. Academic freedom exceptions exist for disclosures made within a class discussion or assignment *related to course content*; under those conditions only, a report to the Title IX Officer is not required.

- The Campus Advocacy Resources and Education (CARE) Office (831) 502-2273, care@ucsc.edu can provide confidential support, resources, and assist with academic accommodations. To make a Title IX report, please contact Tracey Tsugawa, Title IX Officer, (831) 459-2462, ttsugawa@ucsc.edu.

Academic Integrity

- No form of academic dishonesty will be tolerated. You are encouraged to read the campus policies regarding academic integrity at https://www.ue.ucsc.edu/academic_misconduct. Violations may lead to penalties including (but not limited to) failing this course. ***Please be sure to take this seriously.***
- You are allowed to ask for some help when working on assignments, provided that you acknowledge the help that you received on the work that you turn in.
- Points will be deducted if it appears that labor has been divided among multiple students; otherwise, there will be no penalty for small amounts of acknowledged assistance.
- If you have any questions about these rules, please discuss them with the instructor **immediately**.

Disability Resource Center

Special Accommodations:

UC Santa Cruz is committed to creating an academic environment that supports its diverse student body. If you are a student with a disability who requires accommodations to achieve equal access in this course, please submit your Accommodation Authorization Letter from the Disability Resource Center (DRC) to me privately during my office hours or by appointment, preferably within the first two weeks of the quarter.

At that time, I would also like us to discuss ways we can ensure your full participation in the course. I encourage all students who may benefit from learning more about DRC services to contact DRC by phone at 831-459-2089, or by email at drc@ucsc.edu.

Student responsibilities are as follows:

1. Students contact the DRC to determine their eligibility for accommodations. Students will request and receive their accommodation letter. This letter is provided to the instructor. This is official notice of a request for accommodation.
2. Students then notify their instructor during office hours or after class of their accommodations, and provide their instructor with their Accommodation Letters, **preferably during the first two weeks of the term.**
3. Students will manage their own disclosure of disability status, which will be maintained confidentially according to UC Santa Cruz data practices.

What is a Database?

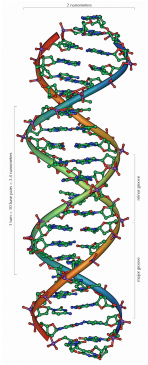
- A *database* is a collection of data.
  - Typically describes the activities of one or more related organizations over time.

- Data Management (including database) is *extremely* important.
  - Essential to every business organization.
  - Enterprises
    - Employee data, sales transactions.
  - Web data
    - Amazon, Twitter, Facebook, IMDB, Google.



Data…is Everywhere

MOST ENTERPRISES TODAY GENERATE
MORE DATA THAN THEY CAN PROCESS



facebook

social

finance

health/
medical

learning

transportation

science

government

retail

entertainment

…AND THE AMOUNT OF DATA IS GROWING AT 50% PER YEAR

- according to IDC

How Much Data is There?

If There Was Already an Ocean of Data in 2007, How Much is there Now?, Andrew MacAfee, MIT

"There were 295 exabytes of digital data in 2007, and (by EMC estimate) 4.4 zettabytes in 2013, giving an annual growth rate of 57.5% over the period."

"If we associate the volume of digital data in the world in 2007 with the volume of the Atlantic Ocean, …

then the volume of datawater created between 2007 and 2013 would cover the Earth to a depth of 84,417 meters (276,000 feet), which is almost ten times the height of Mt. Everest."

From Bytes to Yottabytes

Multiples of bytes					v · t · e
SI decimal prefixes		Binary usage	IEC binary prefixes		
Name (Symbol)	Value		Name (Symbol)	Value	
kilobyte (kB)	10^3	2^{10}	kibibyte (KiB)	2^{10}	
megabyte (MB)	10^6	2^{20}	mebibyte (MiB)	2^{20}	
gigabyte (GB)	10^9	2^{30}	gibibyte (GiB)	2^{30}	
terabyte (TB)	10^{12}	2^{40}	tebibyte (TiB)	2^{40}	
petabyte (PB)	10^{15}	2^{50}	pebibyte (PiB)	2^{50}	
exabyte (EB)	10^{18}	2^{60}	exbibyte (EiB)	2^{60}	
zettabyte (ZB)	10^{21}	2^{70}	zebibyte (ZiB)	2^{70}	
yottabyte (YB)	10^{24}	2^{80}	yobibyte (YiB)	2^{80}	
See also: Multiples of bits · Orders of magnitude of data					

You Should Be Able to Answer These Questions!

Not a homework, but for CMPS180 students, these should be easy.

0-If set S is {1,3,5,7} and set T is {2,3,5,7}, what are S UNION T and S INTERSECT T?

1-If set A is {1,2,3} and set B is {u,v,w,x,y}, how many ways can you pick pairs of items, with the first from A and the second from B?

2-If you have a set of employees (with names and salaries) where John makes 10K, George makes 20K, Ringo makes 30K and Paul makes 40K, what are the names of the employee(s) who make less than the average salary?

3-Can there ever be an employee who makes more than every employee? If so, give an example. If not, explain why not

4-Write the truth-table for p AND q, where p can be TRUE or FALSE and q can be TRUE or FALSE.

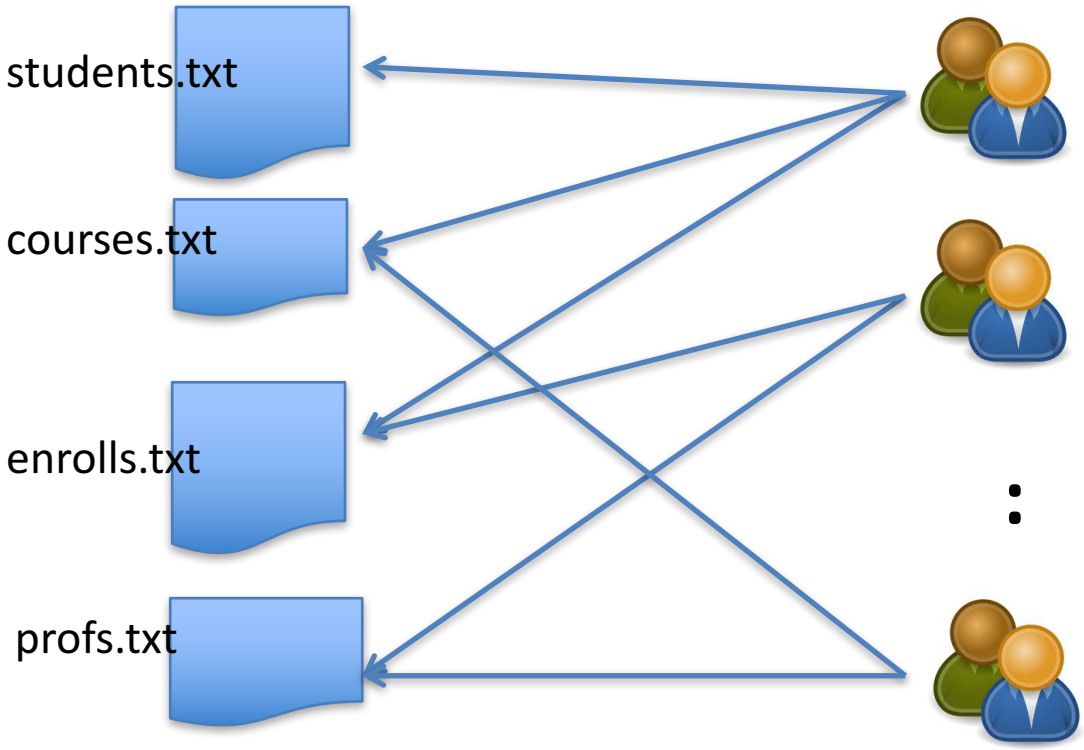
What is a Database Management System?

- A *database management system (DBMS)* is a software system designed to assist in creating, storing, accessing, and updating a database … with enterprise qualities
  - Transactions
  - High/continuous availability
  - Performance
    - Response time/latency
    - Throughput
    - Scalability
  - Security
    - Authorization, access control, etc.

Why Can't we Just Use a File System to Manage our Data?

- Suppose a company has a large database.
  - Data needs to be accessed *frequently* and *concurrently*.
  - Different queries need to *posed easily* be answered *quickly.*
  - Updates to data by different users need to be managed and applied *consistently*.
  - Access to certain parts of the data by certain users need to be *restricted*.
- Another answer, to be discussed later
  - Sometimes we can use file system to manage data
  - … and **much more data** is kept in file systems than in DBs!

Simplified version of MyUCSC Campus Portal on a file system



Did “Ann” get an A for CMPS182 in Fall 2011? Change Bob’s grade to B.

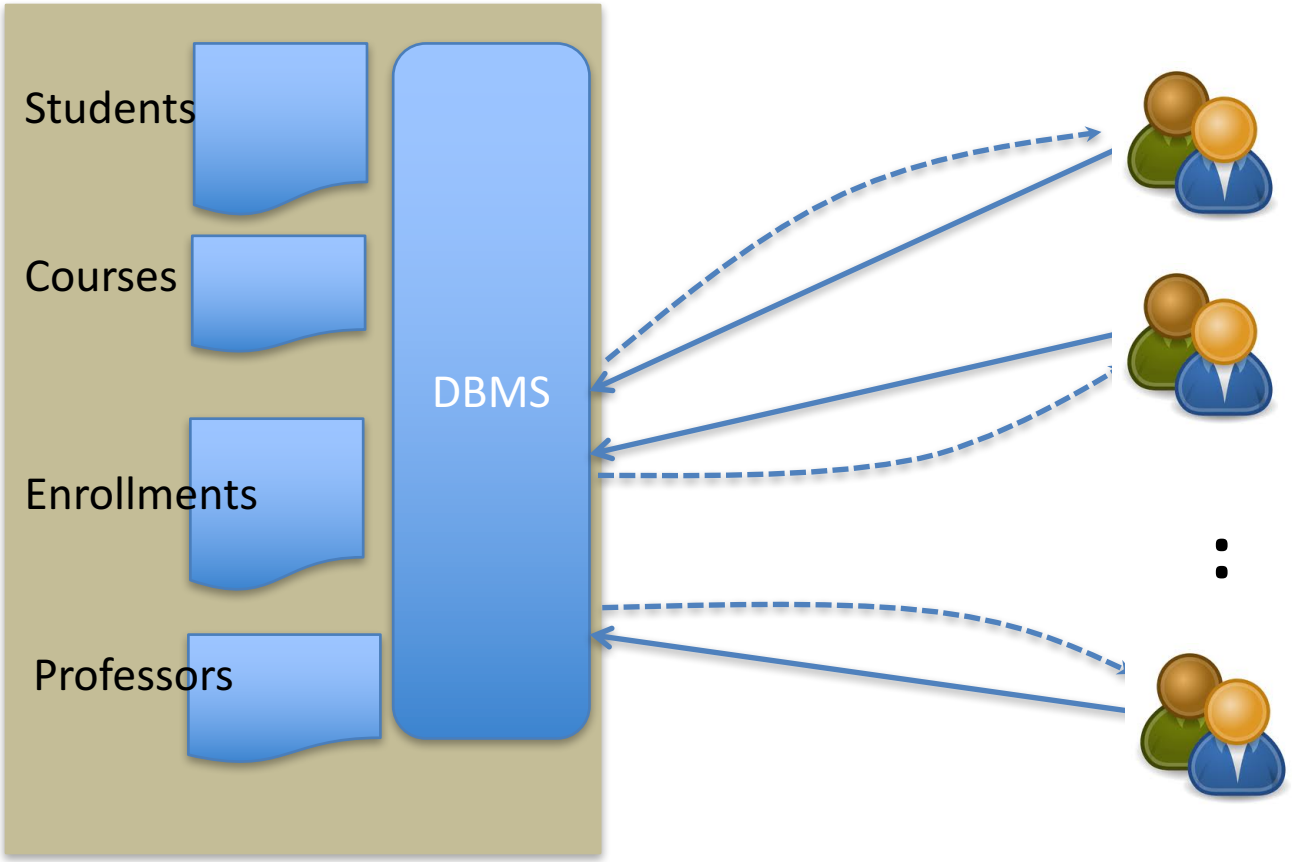
Did “Bob” enroll in a class taught by Prof. Kolaitis? Change Bob’s grade to A+.

Which courses are offered by the CS department in Spring 2011?

Key Characteristics of a DBMS

- Data Model
  - Provides an abstraction of the underlying data.
- High-level language for manipulating data
  - For defining, updating, retrieving and processing data.
- Transaction Processing
  - Concurrent access and updates, crash recovery.
- Access control
  - Limit access of certain data to certain users.

Key Characteristics of a DBMS



Did "Ann" get an A for CMPS182 in Fall 2011? Change Bob's grade to B.

Did "Bob" enroll in a class taught by Prof. Kolaitis? Change Bob's grade to A+.

Which courses are offered by the CS department in Spring 2011?

Advantages of a DBMS

- Users only need to understand the data model and high-level language for manipulating data.
  - Users focuses on *what* data is to be accessed and not *how* data is accessed.
  - Users are not aware of how data is actually stored or laid out on disks.
- Illusion that they are the only users of the DBMS.
- Data integrity is not compromised by system failures.
  - Deposit: balance = balance + 500;
  - In parallel, a withdrawal for your monthly car payment: balance = balance – 300;
  - System crashes… What is the balance?

Advantages of a DBMS (cont'd)

- Queries are automatically optimized for efficiency.
- Integrity of data is automatically enforced.
  - E.g., Employee id is unique, age < 200,
    courses that students are enrolled in must exist.
- Ease of data administration.
  - Well-developed user interfaces.
- Fast application development.
  - Available APIs and libraries.
- Data is managed centrally.
  - Costs are shared across applications.

Transactions have the ACID properties
(to fill in)

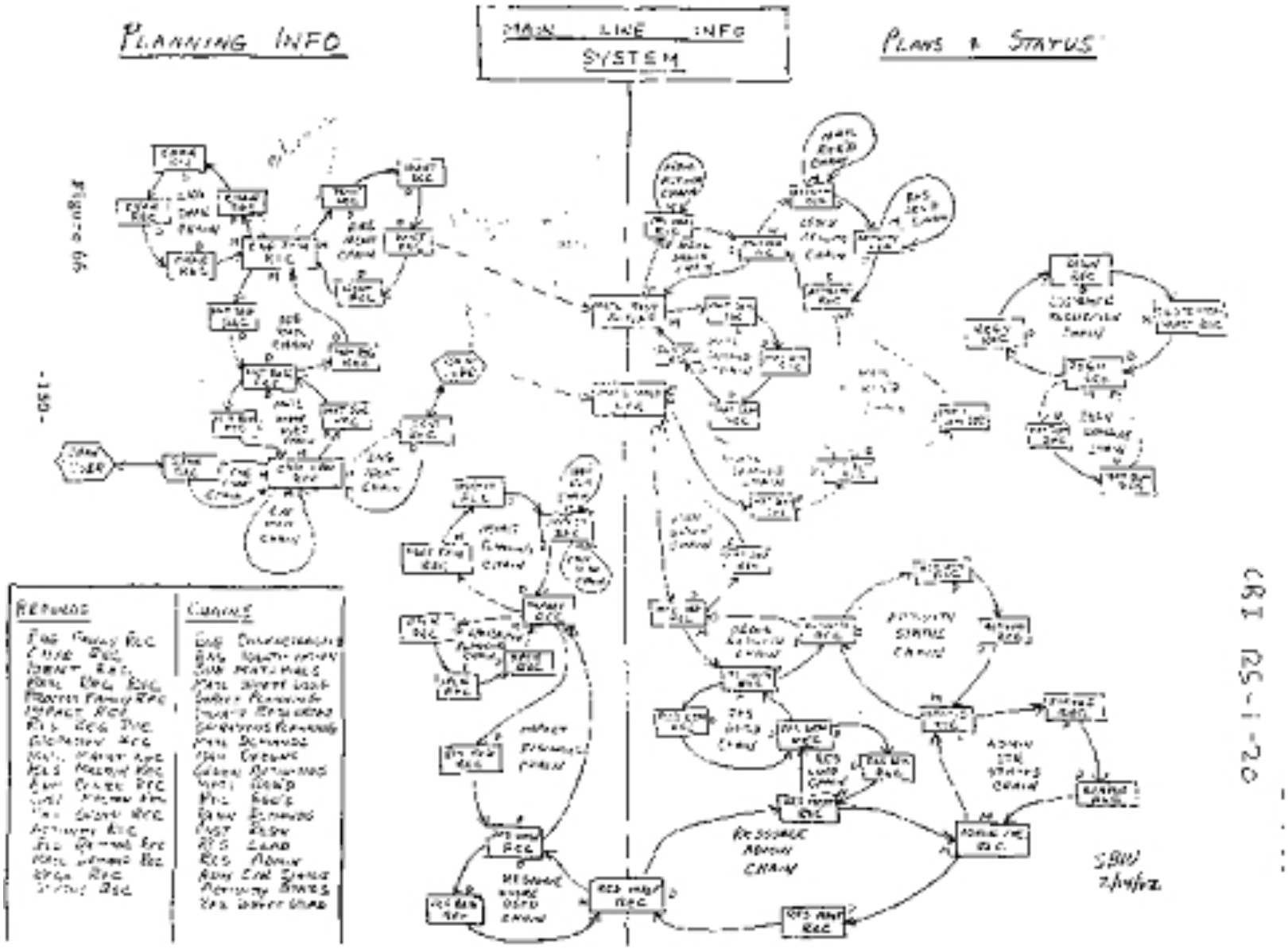
- A(tomicity)
- C(onsistency)
- I(solation)
- D(urability)

But What is a Relational Database Management System (RDBMS)?

- Network Model
- Hierarchical Model
- Relational Model

Network Data Model

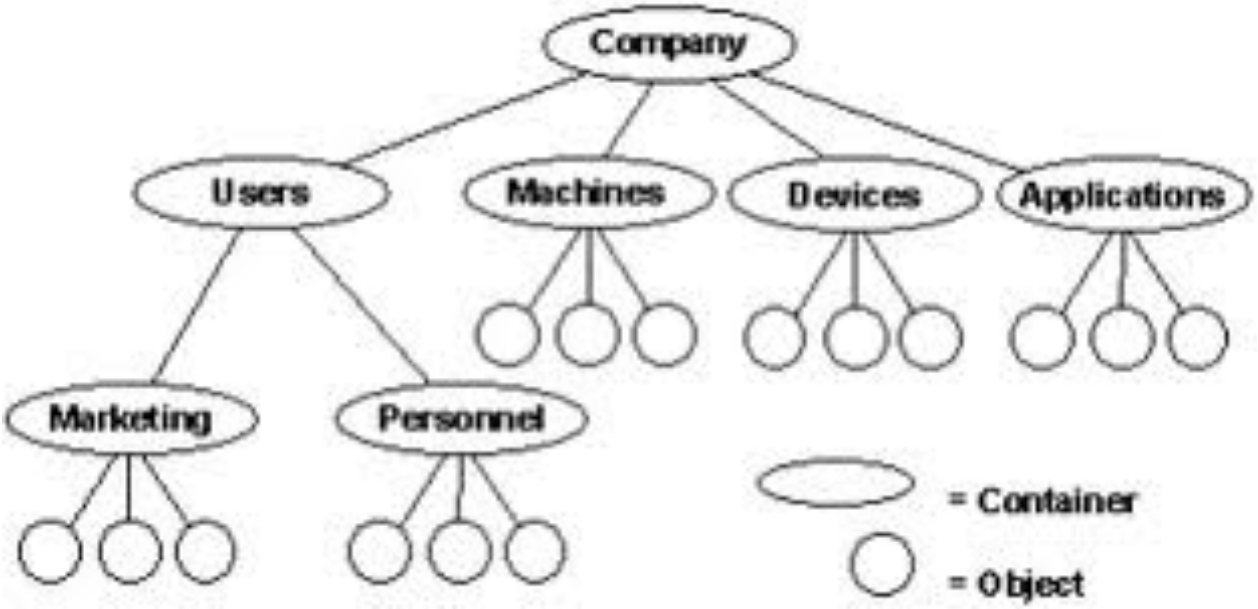
- 1960s
  - First general purpose DBMS was built.
  - Integrated data store (IDS)
    - by Charles Bachman of General Electric.
  - Network Data Model
    - The computer navigates through a space of data records connected by pointers. A graph-based data structure.
    - A user needs to formulate the process of navigating through records and pointers to compute an answer for a query.
  - 1973 Turing award lecture.
    - “The Programmer as Navigator”



Source: Fifty years of databases http://wp.sigmod.org/?p=688

Hierarchical Data Model

- Also in the 1960s:
  - Hierarchical Data Model proposed IBM (IMS product)
  - A tree-based data structure.



Relational Data Model

- 1970s
  - The beginning of *relational* database management systems.
  - Edgar (Ted) F. Codd at the IBM San Jose Research Laboratory (now IBM Almaden Research Center) published a seminal paper:

  "*A relational model for data for large shared data banks*" Communications of the ACM, 1970.

Ted Codd's Relational Model

- Advocates the a radically different data model, called the *relational* data model.
  - *All* data must be stored in flat, table-like relations.
    - No pointers, no hierarchy!
  - Two database query languages:
    - Relational algebra and relational calculus

Employees

EmpNo	FirstName	LastName	Department
100	Sally	Baker	10-L
101	Jack	Douglas	10-L
102	Sarah	Schultz	20-B
103	David	Drachmeier	20-B

Equipment

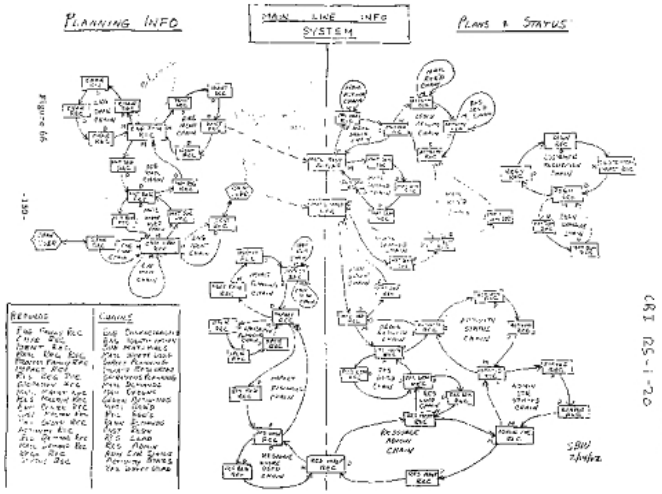
SerialNum	Type	UserEmpNo
3009734-4	Computer	100
3-23-283742	Monitor	100
2-22-723423	Monitor	101
232342	Printer	100

Some Relational Projects and Products

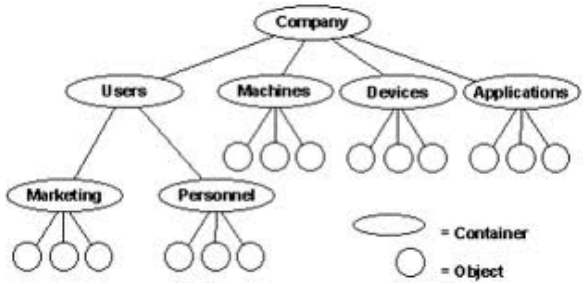
- System R project started at IBM Research in 1974.
  - System R became today's DB2.
  - 1981 Turing Award to Edgar F. Codd, "Relational Database: A Practical Foundation for Productivity".
  - 1998 Turing Award to Jim Gray for Transaction Processing.
- Michael Stonebraker and Eugene Wong at UC Berkeley started INGRES project based on Codd's papers.
  - Relational Technology Inc. became company, INGRES.
  - Later POSTGRES project led to company, Illustra Information Technologies and became open-source PostgreSQL.
  - 2014 Turing Award to Mike Stonebraker.
- Larry Ellison founded Relational Storage Inc., which became Oracle. First Oracle RDBMS was released in 1979 (and was called v2).

RDBMS Today

- Lots of relational database management systems
  - http://en.wikipedia.org/wiki/List_of_relational_database_management_systems
- Examples of open-source RDBMS:
  - MySQL, PostgreSQL
- Examples of proprietary RDBMS:
  - Oracle, IBM DB2, Microsoft SQL Server, SAP HANA



1960s: Network data model and hierarchical data model



1970s: Relational data model

EmpNo	First Name	Last Name	Dept. Num	Serial Num	Type	User EmpNo
100	Sally	Baker	10-L	3009734-4	Computer	100
101	Jack	Douglas	10-L	3-23-283742	Monitor	100
102	Sarah	Schultz	20-B	2-22-723423	Monitor	100
103	David	Drachmeier	20-B	232342	Printer	100


```
{
    "photos": {
        "page": 1,
        "pages": 94276,
        "perpage": 15,
        "total": "1414129",
        "photo": [{
            "id": "3891667770",
            "owner": "35468133120@N01",
            "secret": "4479baabf9",
            "server": "2451",
            "farm": 3,
            "title": "Mexican train dominoes with Brian and Michelle",
            "ispublic": 1,
            "isfriend": 0,
            "isfamily": 0
        },
        {
            "id": "3891661852",
            "owner": "10640301@N07",
            "secret": "79de502257",
            "server": "2590",
            "farm": 3,
            "title": "Peaches",
            "ispublic": 1,
            "isfriend": 0,
            "isfamily": 0
        },
```


2010s: JSON


```
<Books>
    <Book ISBN="0553212419">
        <title>Sherlock Holmes: Complete Novels...
        <author>Sir Arthur Conan Doyle</author>
    </Book>
    <Book ISBN="0743273567">
        <title>The Great Gatsby</title>
        <author>F. Scott Fitzgerald</author>
    </Book>
    <Book ISBN="0684826976">
        <title>Undaunted Courage</title>
        <author>Stephen E. Ambrose</author>
    </Book>
    <Book ISBN="0743203178">
        <title>Nothing Like It In the World</title>
        <author>Stephen E. Ambrose</author>
    </Book>
</Books>
```


1990s: XML (Semi-structured data model)

Today, Data Also Resides Outside Databases

- Before the Web
  - Data typically reside in enterprises, in databases
- Today
  - Data resides in enterprises in on the Web
  - Enterprise data
    - Typically sensitive information.
      - Bank accounts, employee data, sale transactions
  - Data on the Web
    - Amazon, Twitter, Facebook, IMDB, Google, …



The Data Deluge

"… mankind created 150 exabytes (billion gigabytes) of data in 2005. "

– The Data Deluge. The Economist. Feb 2010.

"In 2011 alone, 1.8 zettabytes (or 1.8 trillion gigabytes) of data will be created, the equivalent to every U.S. citizen writing 3 tweets per minute for 26,976 years. And over the next decade, the number of servers managing the world's data stores will grow by ten times."

– IDC study, 2011.

From Bytes to Yottabytes

Multiples of bytes					v · t · e
SI decimal prefixes		Binary usage	IEC binary prefixes		
Name (Symbol)	Value		Name (Symbol)	Value	
kilobyte (kB)	10^3	2^{10}	kibibyte (KiB)	2^{10}	
megabyte (MB)	10^6	2^{20}	mebibyte (MiB)	2^{20}	
gigabyte (GB)	10^9	2^{30}	gibibyte (GiB)	2^{30}	
terabyte (TB)	10^{12}	2^{40}	tebibyte (TiB)	2^{40}	
petabyte (PB)	10^{15}	2^{50}	pebibyte (PiB)	2^{50}	
exabyte (EB)	10^{18}	2^{60}	exbibyte (EiB)	2^{60}	
zettabyte (ZB)	10^{21}	2^{70}	zebibyte (ZiB)	2^{70}	
yottabyte (YB)	10^{24}	2^{80}	yobibyte (YiB)	2^{80}	
See also: Multiples of bits · Orders of magnitude of data					

IDC Digital Universe study. 2011

Today

- NOSQL ("Not Only SQL") systems
  - Map/Reduce (Hadoop, Spark)
  - Column store (HBase, Cassandra)
  - Graph databases (Neo4J, Virtuoso)
  - Document databases (MongoDB)

- Simplicity of design, easy scale-out
- Compromise consistency in favor of availability and partition tolerance
- Lack of full ACID support, use of low-level query languages, lack of standardized interfaces (changing)

Some Supplementary Reading Material

- Database management systems:
  http://en.wikipedia.org/wiki/Database_management_system

- Fifty years of databases:
  http://wp.sigmod.org/?p=688

- The Data Deluge. The Economist.
  http://www.economist.com/node/15579717

- Data, data everywhere. The Economist.
  http://www.economist.com/node/15557443

So … what is a Relational Database Management System (RDBMS)?

RDBMS Components

- Upper half of system: Similar to Language Processing
  - Parsing SQL
  - Query Optimization
  - Plan Generation
  - Security
- Lower half of system: Similar to Operating Systems
  - Storage
  - Plan Execution
  - Concurrency Control and Scheduling
  - Logging and Recovery

DBMS Structure (CMPS 181)