Background

What is a Database?

- A *database* is a collection of data.
  - Typically describes the activities of one or more related organizations over time.
- Databases are *extremely* important.
  - Essential to every business organization.
  - Enterprises
    - Employee data, sales transactions.
  - Web data
    - Amazon, Twitter, Facebook, IMDB, Google, snapchat...



What is a Database?

- A *database* is a collection of data.
  - Typically describes the activities of one or more related organizations over time.
- Databases are *extremely* in[illegible]tant.
  - Essential to every busine[illegible]ganization.
  - Enterprises
    - Employee data, sa[illegible]nsactions, bank accounts.
  - Web data
    - Amazon, T[illegible] Facebook, IMDB, Google, snapchat….

Databases are ubiquitous !

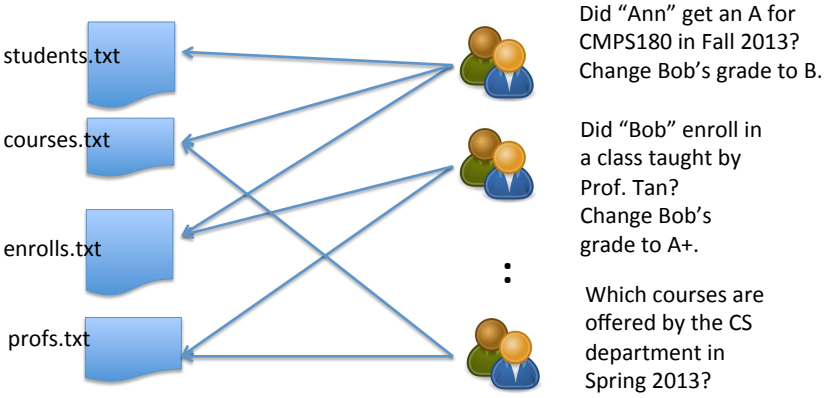
What is a Database Management System?

- A *database management system (DBMS)* is a software designed to assist in creating, storing, accessing, and updating a database.

Why can't we 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*.

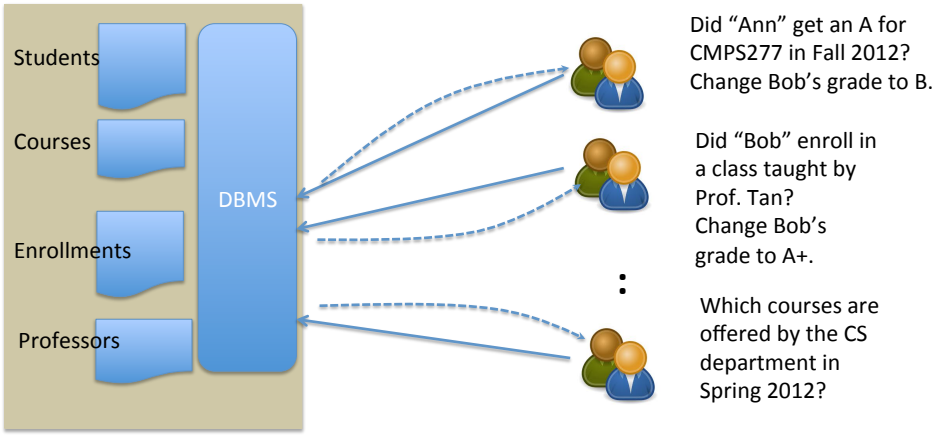
Simplified version of MyUCSC Campus Portal on a file system



Why can't we use a file system to manage our data? (cont'd)

- Special routines will be needed to support the above functionalities over a file system.
  - Data needs to be accessed *frequently* and *concurrently*.
    - Add routines to support efficient and concurrent access.
  - Queries need to be *posed easily* be answered *quickly.*
    - Custom routines are needed for different types of queries.
  - Updates to data by different users need to be managed and applied *consistently*.
    - Add routines to support concurrent access and crash recovery.
  - Access to certain parts of the data by certain users need to be *restricted*.
    - Add routines to enforce access policies.

Key characteristics of a DBMS



Key characteristics of a DBMS

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

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.
- Ease of data administration.
  - Well-developed user interfaces.
- Fast application development.
  - Available APIs and libraries.
- Data is managed centrally.
  - Costs are shared across applications.

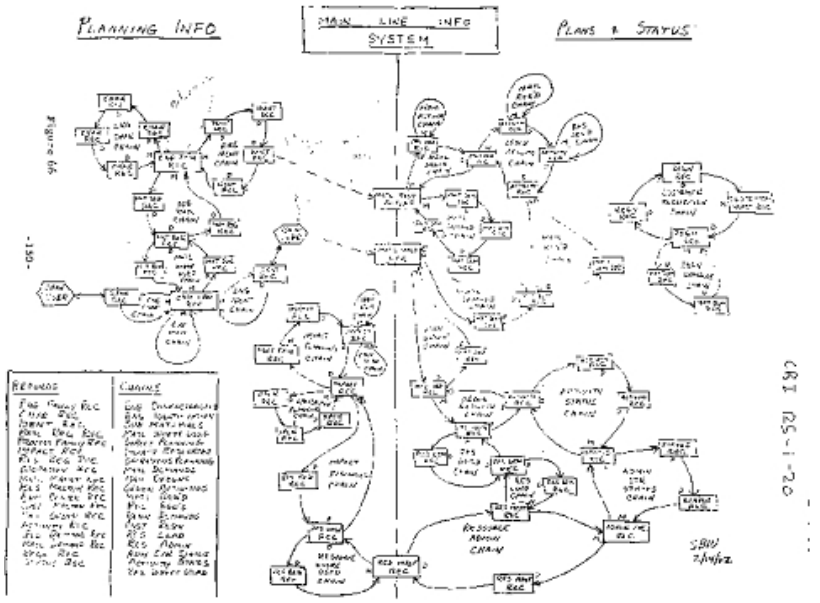
Transactions have the ACID properties

- A(tomicity)
  - All-or-nothing execution of transactions.
    - Transactions, once started, will either be completed or rolled back.
- C(onsistency)
  - Transactions preserve the consistency of constraints of data in the database.
    - E.g., credit limit cannot be negative after a transaction.
- I(solation)
  - Each transaction executes as if no other transactions are executing simultaneously.
- D(urability)
  - The effect of a transaction on the database must never be lost, once the transaction completes.

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

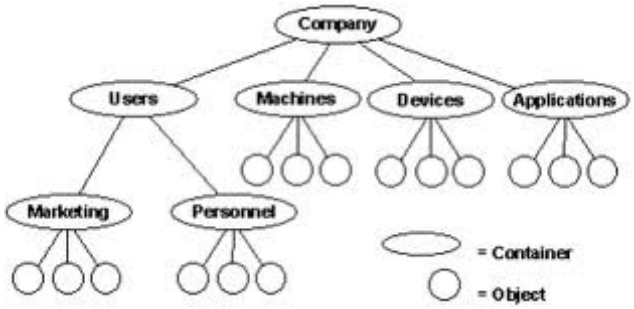
A bit of history

- 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"

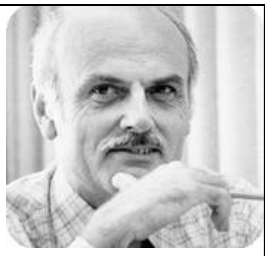


Source: Fifty years of databases http://wp.sigmod.org/?p=688

- Also in the 1960s:
  - Hierarchical Data Model proposed by IBM.
  - A tree-based data structure.



A bit of history (cont'd)



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

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

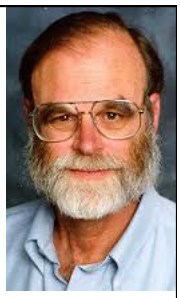
In piazza

- Advocates 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.

EmpNo	First Name	Last Name	Dept. Num
100	Sally	Baker	10-L
101	Jack	Douglas	10-L
102	Sarah	Schultz	20-B
103	David	Drachmeier	20-B

Serial Num	Type	User EmpNo
3009734-4	Computer	100
3-23-283742	Monitor	100
2-22-723423	Monitor	100
232342	Printer	100

- System R project started at IBM San Jose Labs in 1974.
- System R eventually became today's DB2.
- 1981 Turing Award Lecture:
  - "Relational Database: A Practical Foundation for Productivity".
- Michael Stonebraker and Eugene Wong at UC Berkeley started the INGRES project based on Codd's papers.
  - Evolved into Postgres (Post Ingres).
  - Evolved into today's open-source PostgreSQL.
- Larry Ellison founded what is today's Oracle Corporation. First Oracle RDBMS was released in 1979.

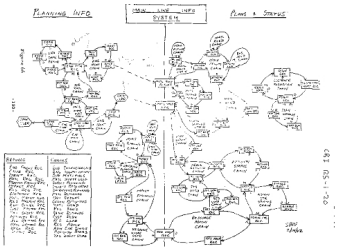


- Jim Gray
  - played a major role in System R
  - created a unified approach to the interrelated problems of concurrency control and crash recovery.
  - developed techniques that allowed concurrent execution of many transactions, as well as restart after crashes, while maintaining the consistency of the database.
  - proved the correctness of the approach.
  - this work led to his Turing Award in 1998.

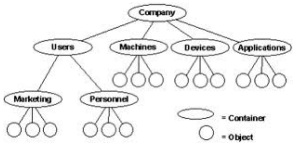
- Michael Stonebraker (Turing award: 2014)
  - the inventor of many concepts that were crucial to making databases a reality and that are used in almost all modern database systems.
  - Introduced the notion of query modification, used for integrity constraints and views.
  - introduced the object-relational model, effectively merging databases with abstract data types while keeping the database separate from the programming language.
  - he released these systems as open software, which allowed their widespread adoption and their code bases have been incorporated into many modern database systems.
  - other influential ideas: implementation techniques for column stores and scientific databases and for supporting on-line transaction processing and stream processing.

RDBMS Today

- Lots of relational database management systems.
- http://en.wikipedia.org/wiki/List_of_relational_database_management_systems
- Examples of open-source relational database management system:
  - MySQL, PostgreSQL



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

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 reside outside enterprises

- Before the Web (and times of Google)
  - Data typically reside in enterprises.
- 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.

amazon.com IMDb Google

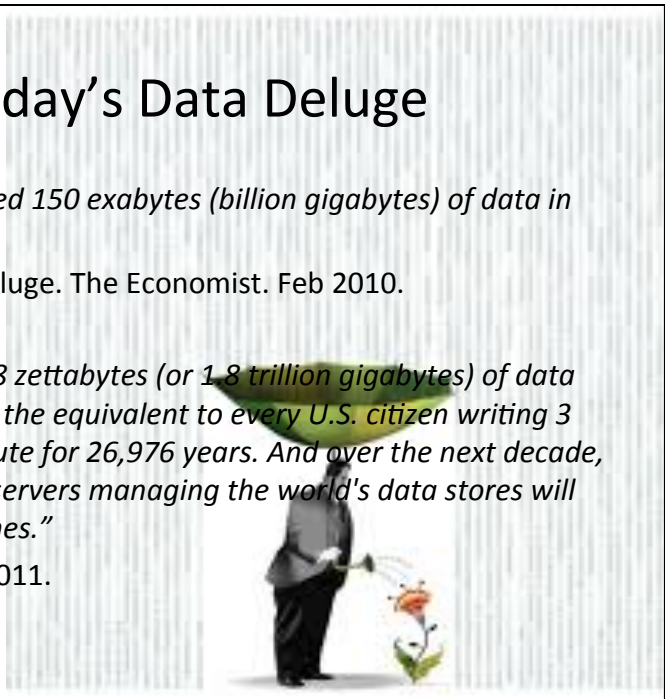
Today's Data Deluge

"… 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

NoSQL databases

- Next generation database systems
  - Handle massive amounts of (non-)relational and schema-free data.
  - Loose model of consistency – "eventual consistency" (not ACID).
  - Distributed, horizontally scalable.
  - NoSQL – "Not Only SQL"
- Popular NoSQL databases
  - Hadoop / Hbase
  - MongoDB
  - Dynamo DB
  - Neo4J

Supplementary reading material

- Database management systems http://en.wikipedia.org/wiki/Database_management_system
- Fifty years of databases http://wp.sigmod.org/?p=688
- NoSQL systems http://nosql-database.org/
- The Data Deluge. The Economist. http://www.economist.com/node/15579717
- Data, data everywhere. The Economist. http://www.economist.com/node/15557443

Research in data management

- Major conferences in data management
  - SIGMOD/PODS, VLDB, ICDE, EDBT/ICDT, …
  - PODS and ICDT are the "theory" conferences
- Major journals in data management
  - ACM TODS, VLDB J, IEEE TKDE, …
- Top data management research groups in industrial labs:
  - IBM Research – Almaden (and several other locations)
  - Microsoft Research – Redmond, ~~Silicon Valley~~
  - AT&T Research Labs
  - HP Labs
  - ~~Yahoo! Research~~

END