

Welcome to

CS 460: Introduction to Database Systems

<https://midas.bu.edu/classes/CS460/>

Instructor: *Manos Athanassoulis*

email: mathan@bu.edu

Today

big data

data-driven world

databases & database systems



when you see this, I want you to
speak up!
[and you can always interrupt me]



no smartphones



no laptop

Big Data

marketing term ...

but ...

science / government / business / personal data

exponentially growing data collections

So, it is all good!

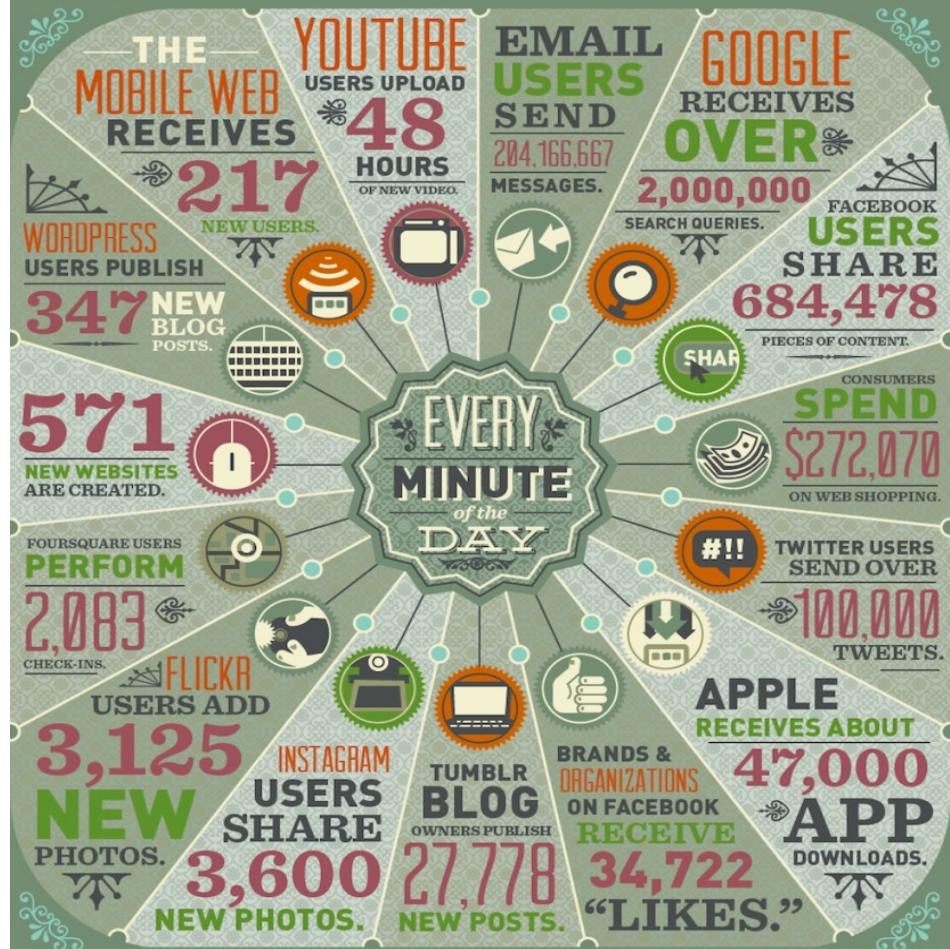
How big is “Big”?



Every day, we create 2.5 exabytes* of data — 90% of the data in the world today has been created in the last two years alone.

[Understanding Big Data, IBM]

*exabyte = 10^9 GB



WITH NO SIGNS OF SLOWING, THE DATA KEEPS GROWING

These are just some of the more common ways that Internet users add to the big data pool. In truth, depending on the niche of business you're in, there are virtually countless other sources of relevant data to pay attention to. Consider the following:

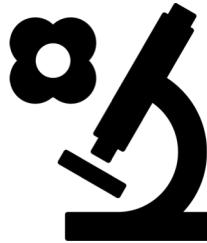
The global Internet population grew 4.57 percent from 2010 to 2011 and now represents 2.1 BILLION PEOPLE.

These users are real, and they are out there leaving data trails everywhere they go. The team at Domo can help you make sense of this seemingly insurmountable heap of data, with solutions that help executives and managers bring all of their critical information together in one intuitive interface, and then use that insight to transform the way they run their business. To learn more, visit www.domo.com.

SOURCES: [HTTP://NEWS.INVESTORS.COM](http://NEWS.INVESTORS.COM), ROYAL.PINGD.COM, BLOG.GROVO.COM, BLOG.HUBSPOT.COM, SIMPLYESTY.COM, PCWORLD.COM, BIZTECHMAGAZINE.COM, DIGBY.COM



Using Big Data



*experimental physics (IceCube, CERN)
biology
neuroscience*



*data mining business datasets
machine learning for corporate and consumer*



data analysis for fighting crime

... are only some examples

Data-Driven World



Big Data V's

Volume

Velocity

Variety

Veracity

Information is transforming traditional business.

[“Data, data everywhere”, Economist]

Data-Driven World

Reporting

Discovery

Logging

Exploration

Transactions

Data-to-Insight

Business Analysis

Automated Decisions

*Behind all these: use &
manage data*

CS460

we live in a ***data-driven*** world

CS460 is about the ***basics*** for
storing, using, and managing data

your lecturer (that's me!)

Manos Athanassoulis

name in greek: Μάνος Αθανασούλης

grew up in Greece

enjoys playing basketball and the sea



photo for VISA / conferences



Myrtos, Kefalonia, Greece

BSc and MSc @ University of Athens, Greece

PhD @ EPFL, Switzerland

Research Intern @ IBM Research Watson, NY

Postdoc @ Harvard University

some awards:

SNSF Postdoc Fellowship

IBM PhD Fellowship

Best of SIGMOD 2017, VLDB 2017

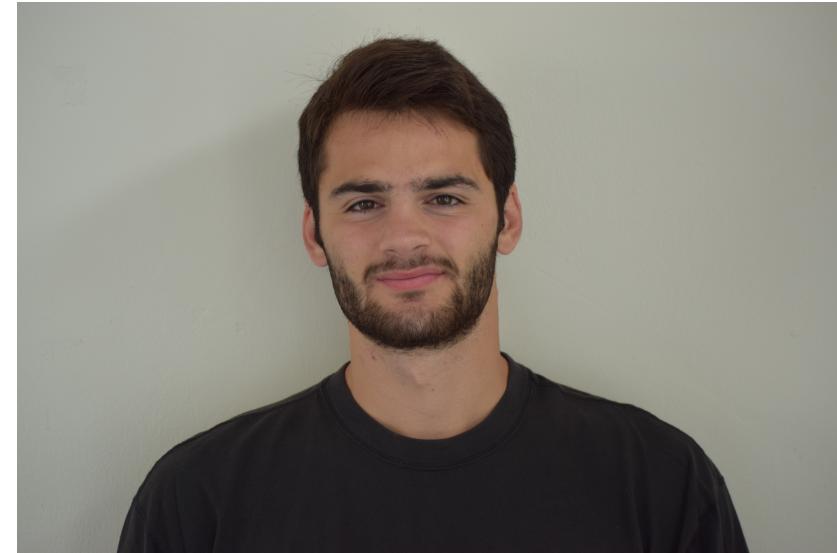
<http://cs-people.bu.edu/mathan/>

Office: MCS 106

Office Hours: M/W before class

your awesome TA

Dimitris Staratzis
grad student in DB



dstara@bu.edu

Data

to make data usable and manageable

we organize them in collections

Databases

a large, integrated, *structured* collection of data

intended to model some real-world enterprise

Examples: a university, a company, social media

University: students, professors, course

what is missing?

-- how to connect these?

-- enrollment, teaching

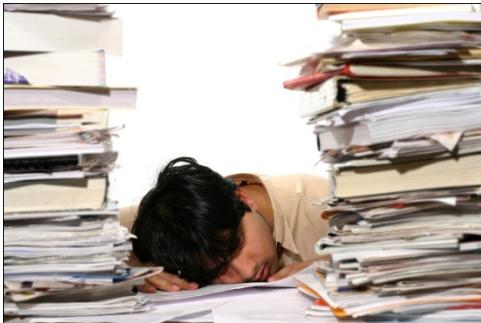


What about a company? What about social media?

Database Systems

a.k.a. database management systems (DBMS)

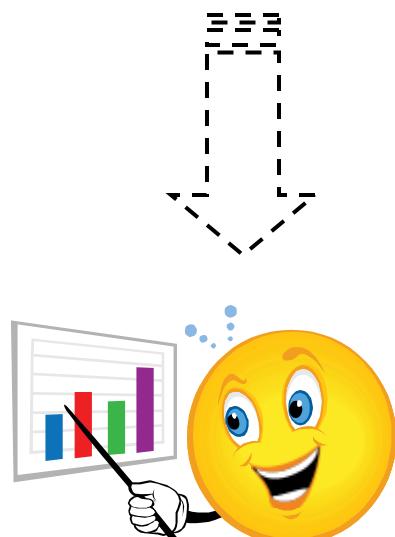
a.k.a. data systems



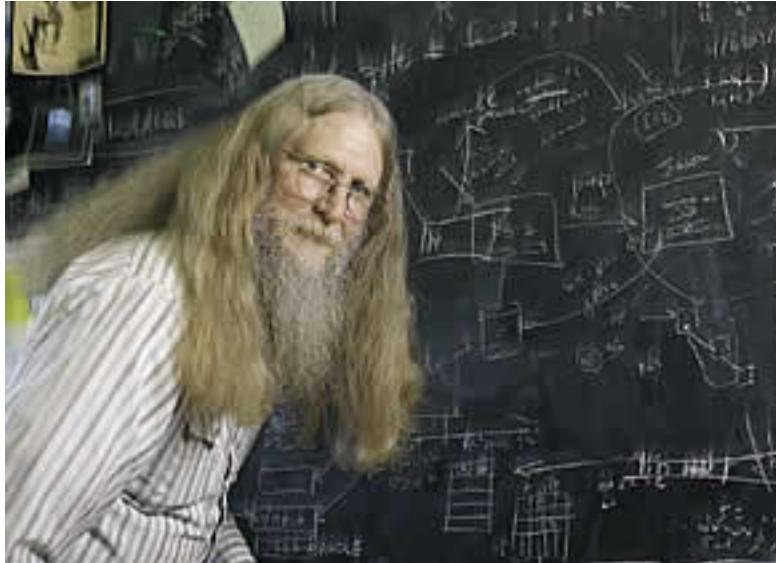
Sophisticated
pieces of software...



... which store, manage,
organize, and facilitate
access to my databases ...



... so I can do things (and ask questions) that are
otherwise hard or impossible



*“relational databases
are the foundation of
western civilization”*

Bruce Lindsay, IBM Research
ACM SIGMOD Edgar F. Codd Innovations award 2012

Ok but what really IS a database system?

Is the WWW a DBMS?



Is a File System a DBMS?



Is Facebook a DBMS?



Is the WWW a DBMS?

Not really!

Fairly sophisticated search available

web crawler *indexes* pages for fast search

.. but

data is unstructured and untyped

not well-defined “correct answer”

cannot update the data

freshness? consistency? fault tolerance?

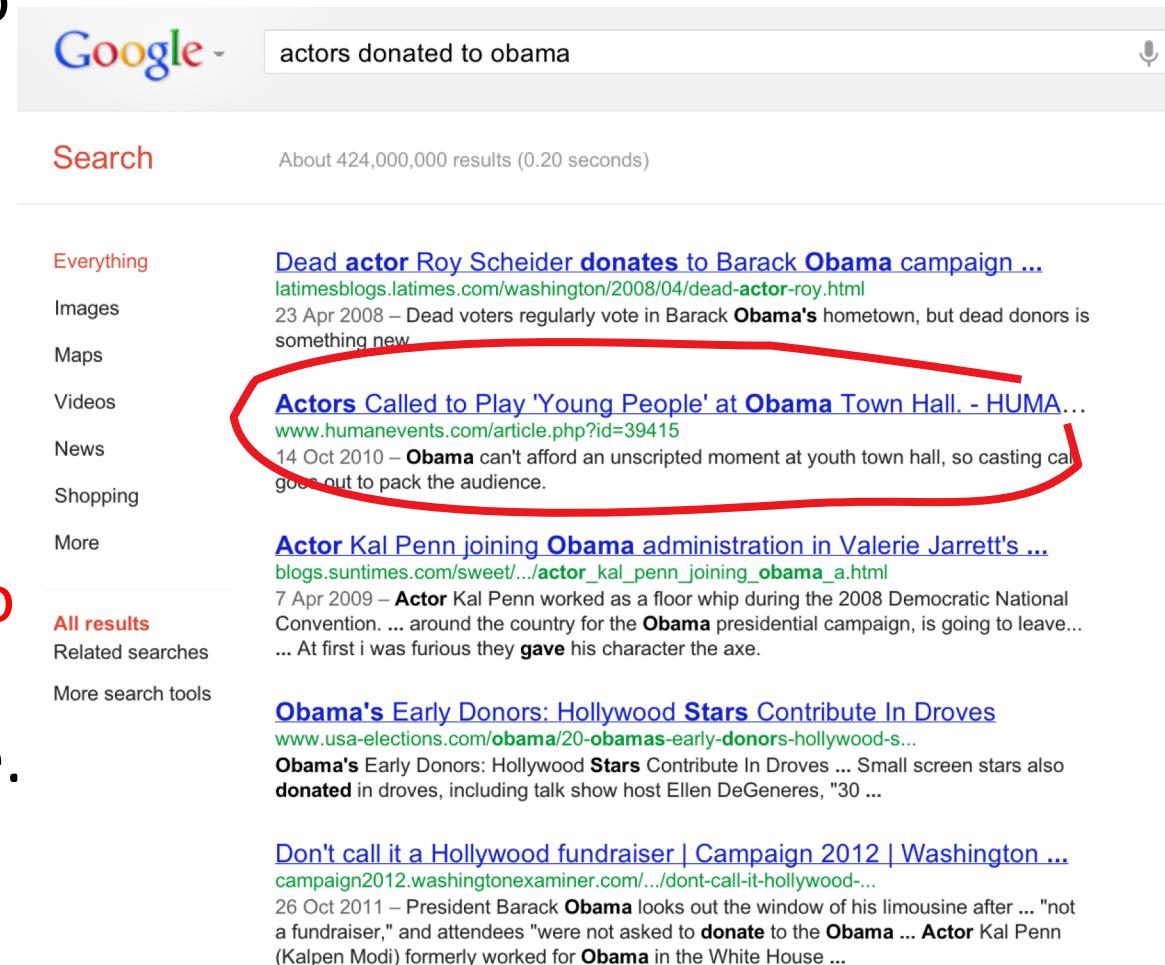
web sites **use** a *DBMS* to provide these functions

e.g., amazon.com (Oracle), facebook.com (MySQL and others)

“Search” vs. Query

What if you wanted to find out which actors donated to the first Barrack Obama's presidential campaign 11 years ago?

Try “actors donated to obama” in your favorite search engine.



Google - actors donated to obama

Search About 424,000,000 results (0.20 seconds)

Everything

Images

Maps

Videos

News

Shopping

More

All results

Related searches

More search tools

Dead actor Roy Scheider donates to Barack Obama campaign ...
latimesblogs.latimes.com/washington/2008/04/dead-actor-roy.html
 23 Apr 2008 – Dead voters regularly vote in Barack **Obama's** hometown, but dead donors is something new

Actors Called to Play 'Young People' at Obama Town Hall. - HUMA...
www.humanevents.com/article.php?id=39415
 14 Oct 2010 – **Obama** can't afford an unscripted moment at youth town hall, so casting call goes out to pack the audience.

Actor Kal Penn joining Obama administration in Valerie Jarrett's ...
blogs.suntimes.com/sweet/.../actor_kal_penn_joining_obama_a.html
 7 Apr 2009 – **Actor** Kal Penn worked as a floor whip during the 2008 Democratic National Convention. ... around the country for the **Obama** presidential campaign, is going to leave...
 ... At first i was furious they **gave** his character the axe.

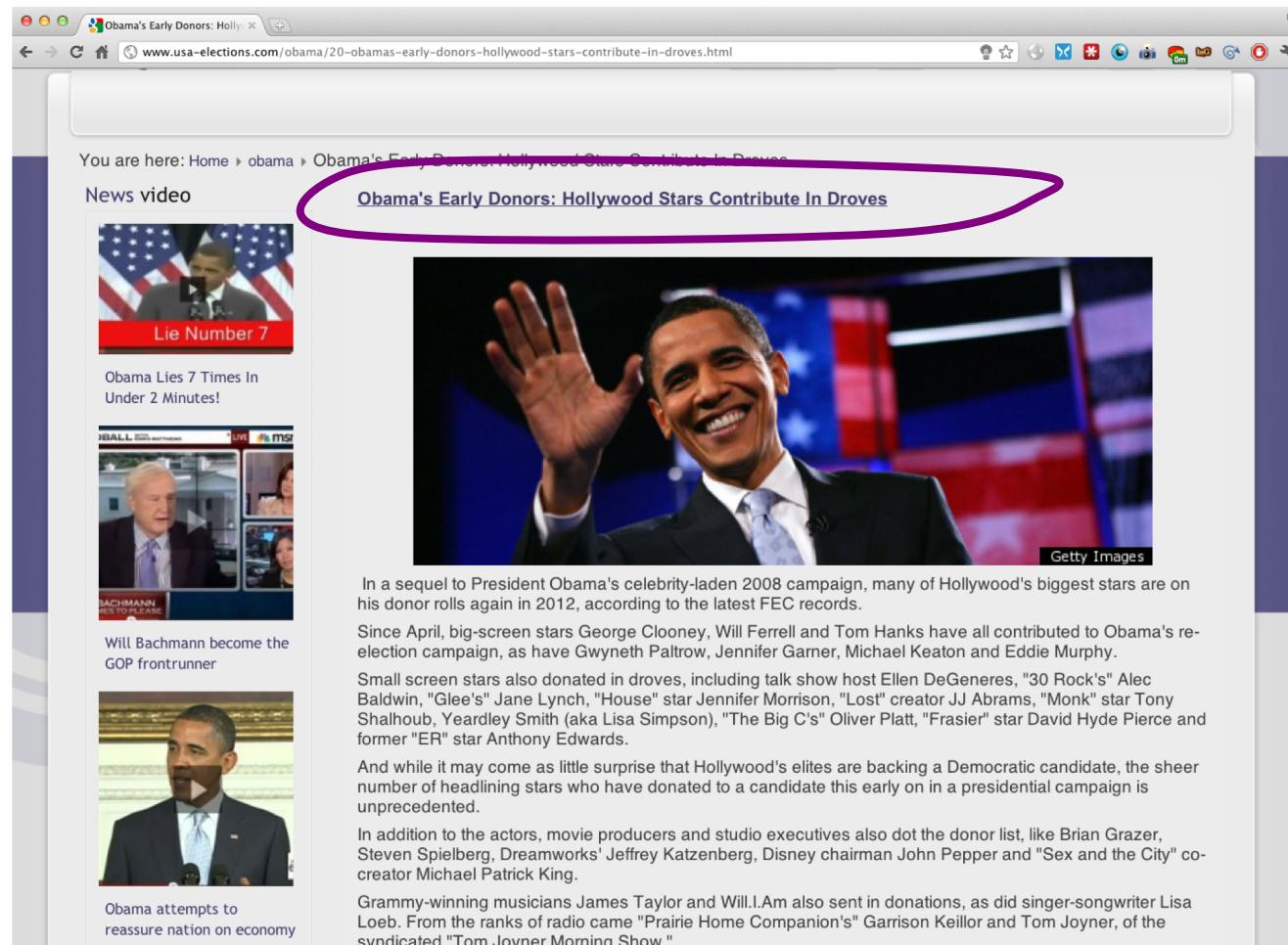
Obama's Early Donors: Hollywood Stars Contribute In Droles
www.usa-elections.com/obama/20-obamas-early-donors-hollywood-s...
Obama's Early Donors: Hollywood **Stars** Contribute In Droles ... Small screen stars also **donated** in droves, including talk show host Ellen DeGeneres, "30 ...

Don't call it a Hollywood fundraiser | Campaign 2012 | Washington ...
campaign2012.washingtonexaminer.com/.../dont-call-it-hollywood-...
 26 Oct 2011 – President Barack **Obama** looks out the window of his limousine after ... "not a fundraiser," and attendees "were not asked to **donate** to the **Obama** ... **Actor** Kal Penn (Kalpen Modi) formerly worked for **Obama** in the White House ...

“Search” vs. Query

“Search” can return only what’s been “stored”

E.g., best match at Google:



A “Database Query” Approach

where can we find
data for “all actors”?



where can we find
data for “all donations”?



A “Database Query” Approach

IMDb Search Results (Left):

- URL: www.imdb.com/search/name?gender=male&sort=alpha,asc&start=16684
- Section: Males
- Sort: Sorted by Name Ascending
- Count: 16,684-16,733 of 1,865,455 names.
- Options: Sort by: STARmeter | A-Z▲ | Height | Birth Date | Death
- Results:
 - 16684.  **Adam Sandler** Producer, *Grown Ups*
Adam Sandler was born on September 9, 1969, in New York City. At 17, he took his first acting job in a school play. He then landed a role in the TV show *Friday Night Lights*. In 2002, he starred in the critically acclaimed film *An Officer and a Gentleman*.
 - 16685.  **Adam Sandler** Producer, Episode #38.2
 - 16686.  **Adam Sandoval** Actor, *Unspeakable*
 - 16687.  **Adam Sandroni** Actor, *Joey's Girl*

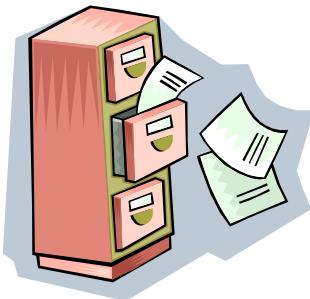
OpenSecrets.org Presidential Donor Lookup Results (Right):

- URL: www.opensecrets.org/pres08/search.php?cid=N00009638&name=%28all%29&employ=%28any+employer%29&state=%28all%29&zip=%2...
- Section: Politicians & Elections > Influence & Lobbying > News & Analysis > Resources > Take Action > About Us > Donate!
- Search Criteria:
 - Sort by Name (radio button)
 - Sort by Date (Descending) (radio button)
 - Sort by Amount (radio button)
 - Sort button
- Message: Your search has generated too many results. Only the top 1000 records are being displayed. If you would like to refine your search, return to the [form page](#).
- Table: Presidential Donor Lookup Results (partial data)

Candidate	Contributor	Employer	Date	Amount
Obama, Barack	Budinger, William Aspen, CO 81611	Not employed	7/31/08	\$30,800
Obama, Barack	BOSLER, JAMES FORT WORTH, TX 76126	NOT EMPLOYED/RETIRED	8/28/08	\$28,500
Obama, Barack	HIGDON, JOE WASHINGTON, DC 20008	NOT EMPLOYED/RETIRED	8/28/08	\$28,500
Obama, Barack	MYERS, DEBRA RANCHO PALOS VERDE, CA 90275	SELF EMPLOYED/PHYSICIAN	8/31/08	\$28,500
Obama, Barack	MYERS, WOODROW DR JR INDIANAPOLIS, IN 46204	MYERS VENTURES LLC/MGR DIRECTOR	8/31/08	\$28,500

“IMDB Actors” JOIN “OpenSecrets”

Contributor	Employer	Date	Amount
ROCK, CHRIS MR NEW YORK,NY 10019	ACTOR	4/20/07	\$9,200
DOUGLAS, MICHAEL UNIVERSAL CITY,CA 91608	ACTOR/ PRODUCER	3/30/07	\$4,600
DOUGLAS, MICHAEL UNIVERSAL CITY,CA 91608	ACTOR/ PRODUCER	3/30/07	\$2,300
ROCK, CHRIS MR NEW YORK,NY 10019	ACTOR	4/20/07	\$2,300
CARIDES, GEORGIA NEW YORK,NY 10017	ACTOR	5/18/07	\$1,000
CARTER COVINGTON, CLAUDIA CHARLOTTE,NC 28207	ACTORS THEATRE PART TIME/ACTOR/NEW	5/20/08	\$1,000
FOX, RICK ENCINO,CA 91316	ACTOR/PRODUCER	6/16/08	\$1,000
HILDRETH, THOMAS W LOS ANGELES,CA 90068	ACTOR	9/29/08	\$1,000
RENNER, CARL BEVERLY HILLS,CA 90210	ACTOR/BESSONE@ROADRUNNER.COM	8/28/08	\$1,000
SIMMONS, HENRY WEST HOLLYWOOD,CA 90046	ACTOR	6/4/07	\$1,000



Is a File System a DBMS?

Not really!

Thought Experiment 1:

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



- A) Yours B) Partner's C) Both D) Neither E) ???**

Thought Experiment 2:

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



- A) All B) None C) All Since last save D) ???**



Is Facebook a DBMS?

Is the data structured & typed?

Does it offer well-defined queries?

Does it offer properties like “durability” and “consistency”?

Facebook is a data-driven company that uses several database systems (>10) for different use-cases (internal or external).



Not really!

Why take this class?

computation to information

corporate, personal (web), science (big data)

database systems *everywhere*

data-driven world, data companies

DBMS: much of CS as a practical discipline

languages, theory, OS, logic, architecture, HW

CS460 in a nutshell

model

data representation model

query

query languages – ad hoc queries

access (concurrently multiple reads/writes)
ensure *transactional* semantics

store (reliably)

maintain *consistency/semantics* in *failures*

A “free taste” of the class

data modeling

query languages

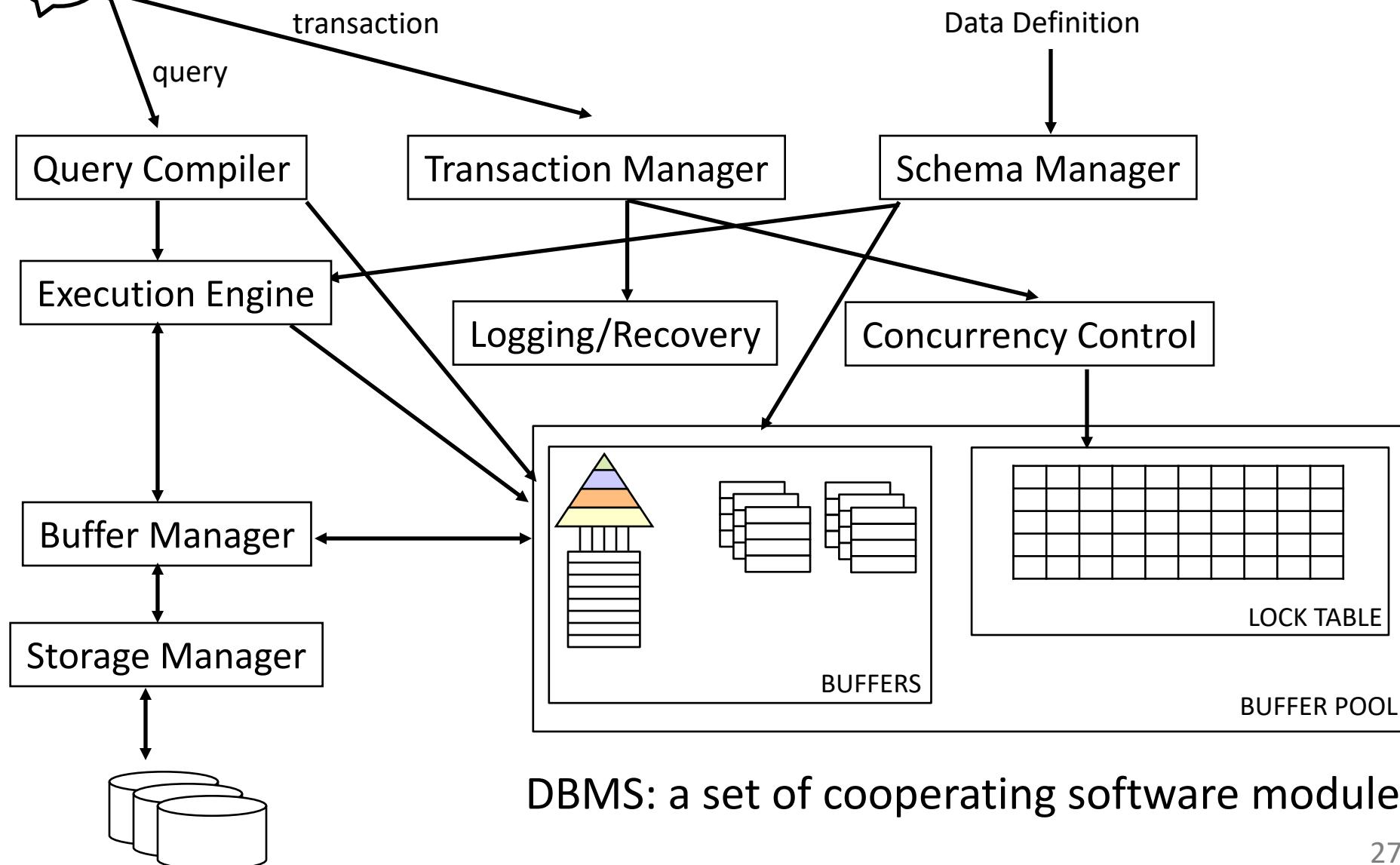
concurrent, fault-tolerant data management

DBMS architecture

Coming in next class

Discussion on *database systems designs*

Components of a “classic” DBMS



Describing Data: Data Models

data model : a collection of concepts describing data

relational model is the most widely used model today
key concepts

relation : basically a table with rows and columns

schema : describes the columns (or fields) of each table

Schema of “University” Database

Students

sid: string, name: string, login: string, age: integer, gpa: real

Courses

cid: string, cname: string, credits: integer

Enrolled

sid: string, cid: string, grade: string



Levels of Abstraction

what the users *see*

External Schema 1

External Schema 2

what is the *data model*

Conceptual Schema

how the data is *physically* stored
e.g., files, indexes

Physical Schema

Schemas of “University” Database

Conceptual Schema

Students

sid: string, name: string, login: string, age: integer, gpa: real

Courses

cid: string, cname: string, credits: integer

Enrolled

sid: string, cid: string, grade: string

Physical Schema

relations stored in heap files

indexes for sid/cid

Schemas of “University” Database

External Schema

a “view” of data that can be derived from the existing data

example: Course Info

Course_Info (cid: string, enrollment:integer)

Data Independence

Abstraction offers “application independence”

Logical data independence

Protection from changes in *logical* structure of data

Physical data independence

Protection from changes in *physical* structure of data

Q: Why is this particularly important for DBMS?

Applications can treat DBMS as
black boxes!



Queries

“Bring me all students with gpa more than 3.0”

“SELECT * FROM Students WHERE gpa>3.0”

SQL – a powerful declarative query language

treats DBMS as a black box

What if we have multiples accesses?

Concurrency Control

multiple users/apps

Challenges



how frequent access to slow medium

how to keep CPU busy

how to avoid *short jobs* waiting behind *long ones*

e.g., ATM withdrawal while summing all balances

interleaving actions of different programs

Concurrency Control

Problems with *interleaving* actions of diff. programs



Bill



Move 100 from
savings to checking



Balance?



Alice

Bad interleaving:

Savings -= 100

Print balances

Checking += 100

Printout is missing 100\$!

Concurrency Control

Problems with *interleaving* actions of diff. programs



Bill



Move 100 from
savings to checking



Balance?



Alice

What is a correct interleaving?

Savings -= 100

Checking += 100

Print balances

How to achieve this interleaving?



Scheduling Transactions

Transactions: atomic sequences of **Reads & Writes**

$$T_{\text{Bill}} = \{R1_{\text{Savings}}, R1_{\text{Checking}}, W1_{\text{Savings}}, W1_{\text{Checking}}\}$$
$$T_{\text{Alice}} = \{R2_{\text{Savings}}, R2_{\text{Checking}}\}$$

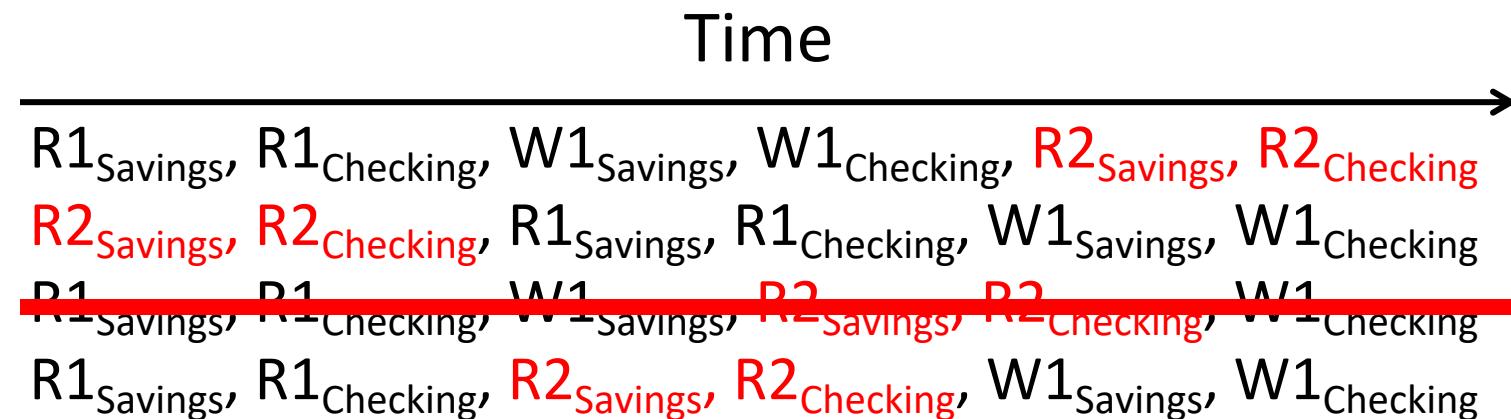
How to avoid previous problems?



Scheduling Transactions

All interleaved executions equivalent to a serial

All actions of a transaction executed as a whole



How to achieve one of these?



Locking



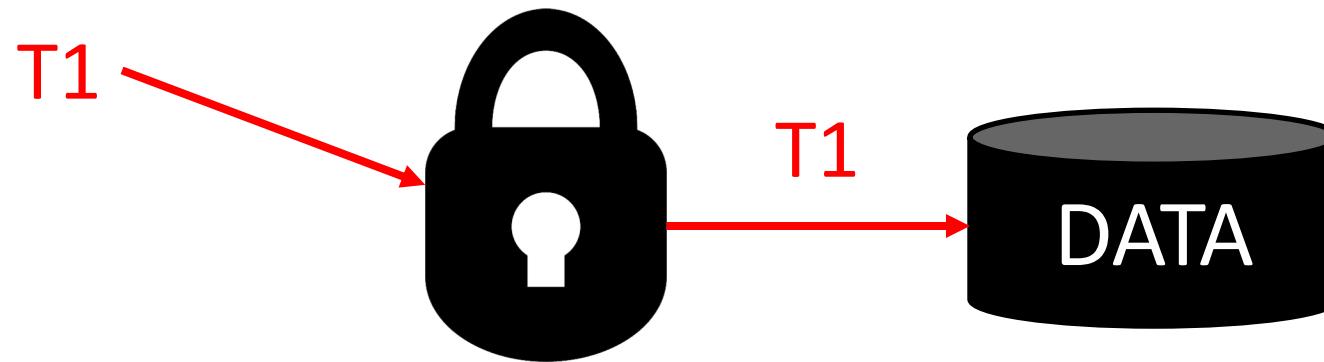
before an object is accessed a lock is requested

Locking



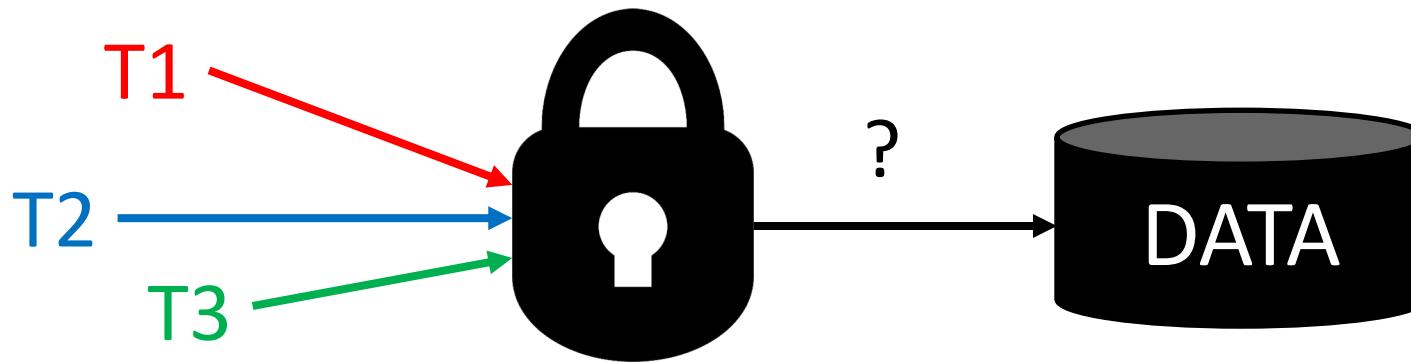
before an object is accessed a lock is requested

Locking



before an object is accessed a lock is requested

Locking



locks are held until the end of the transaction

*[this is only one way to do this, called
“strict two-phase locking”]*

Locking

$T_1 = \{R1_{Savings}, R1_{Checking}, W1_{Savings}, W1_{Checking}\}$

$T_2 = \{R2_{Savings}, R2_{Checking}\}$

Both should lock *Savings* and *Checking*

What happens:

if T1 locks Savings & Checking ?

T2 has to wait

if T1 locks Savings & T2 locks Checking ?

we have a deadlock



How to solve deadlocks?

we need a mechanism to undo

also when a transaction is incomplete

e.g., due to a crash



what can be an undo mechanism?



log every action before it is applied!

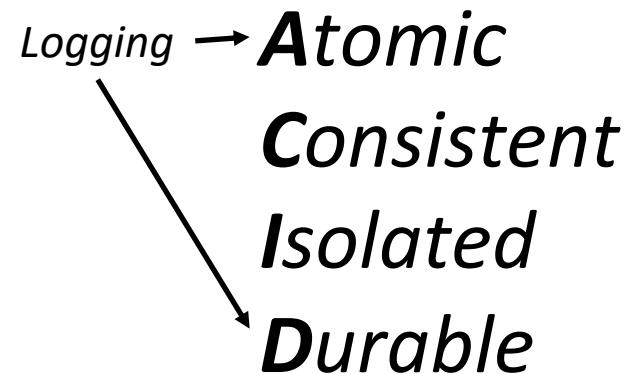
Transactional Semantics

Transaction: one execution of a user program

multiple executions → multiple transactions

Every transaction:

Logging → ***Atomic***
Consistent
Isolated
Durable

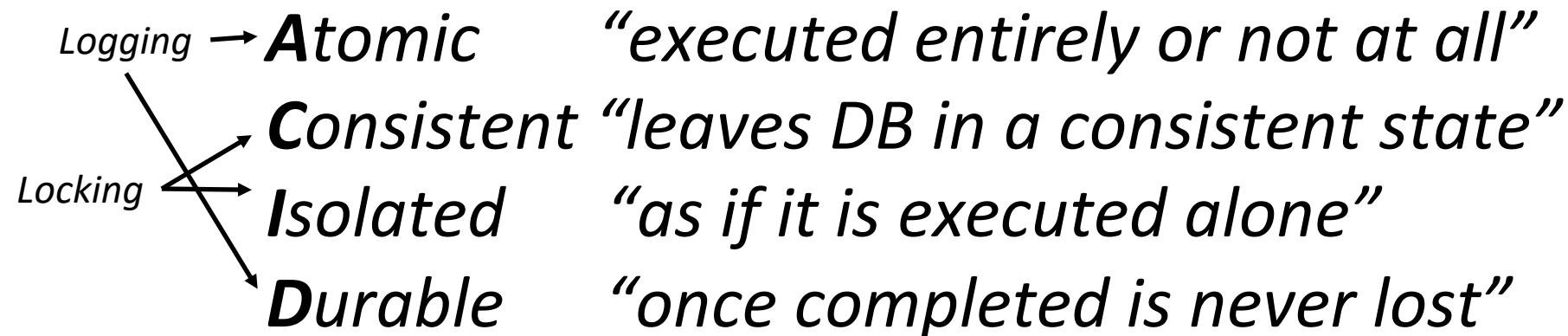


Transactional Semantics

Transaction: one execution of a user program

multiple executions → multiple transactions

Every transaction:



Who else needs transactions?



lots of data

lots of users

frequent updates

background game analytics

Scaling games to epic proportions,

by W. White, A. Demers, C. Koch, J. Gehrke and R. Rajagopalan

ACM SIGMOD International Conference on Management of Data, 2007

Only “classic” DBMS?

No, there is much more!

NoSQL & Key-Value Stores: No transactions, focus on queries

Graph Stores

Querying raw data without loading/integrating costs

Database queries in large datacenters

New hardware and storage devices

... many exciting open problems!

<https://midas.bu.edu/classes/CS460/>

Next time in ...

CS 460: Introduction to Database Systems

Database Systems Architectures

Class administrivia

Class project administrivia

<https://midas.bu.edu/classes/CS460/>

Additional Accommodations

If you require additional accommodations please contact the Disability & Access Services office at aslods@bu.edu or 617-353-3658 to make an appointment with a DAS representative to determine which are the appropriate accommodations for your case.

Please be aware that accommodations cannot be enacted retroactively, making timeliness a critical aspect for their provision.

You can optionally choose to disclose this information to the instructor.