Administrivia

http://db.csail.mit.edu/6.830
Email: 6.830-staff@mit.edu
Ask questions on Piazza!

Lecturers:
   Tim Kraska <kraska@mit.edu>
   Raul Castro Fernandez <raulcf@mit.edu>

TAs:
   Famien Koko <fakoko@mit.edu>
   Nayoung Lee <nayoung@mit.edu>

Guest Lectures:
   Turing-award winner Mike Stonebraker
   The Dave DeWitt (the reason why every commercial DB has a DeWitt clause)
   Richard Hipp, the SQL-Lite guy
   and more
TA Office Hours

• Monday 4-5pm
• Wednesday 4-5pm
Topics

- How to use a Data Management System
  - Data Modeling
  - Relational Algebra
  - SQL
  - Transactions
  - ...
- How Data Management Systems are built
  - Query Operators
  - Query Optimizer
  - Concurrency Control
  - Storage formats
  - Recovery
  - Distributed data management systems
  - ...
- Alternative Systems and Other Data Management Issues
  - Key/Value Stores
  - Data Integration
  - Streaming Systems
  - ...
- New: **For Each Topic We Will Highlight Current Research Efforts**
<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>W1</td>
<td>Sep. 5</td>
<td>Lec 1: Intro to databases</td>
</tr>
<tr>
<td>W2</td>
<td>Sep. 10</td>
<td>Lec 2: The Relational Model</td>
</tr>
<tr>
<td>W2</td>
<td>Sep. 12</td>
<td>Lec 3: Schema Design</td>
</tr>
<tr>
<td>W3</td>
<td>Sep. 17</td>
<td>Lec 4: Intro to DB Internals</td>
</tr>
<tr>
<td>W3</td>
<td>Sep. 19</td>
<td>Lec 5: DB Operators and Query Proc</td>
</tr>
<tr>
<td>W4</td>
<td>Sep. 24</td>
<td>Lec 6: Indexing and Access Methods</td>
</tr>
<tr>
<td>W4</td>
<td>Sep. 26</td>
<td>Lec 7: DB Layout for Analytic DBs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Due: Final Project Teams (6.830 Only)</td>
</tr>
<tr>
<td>W5</td>
<td>Oct. 1</td>
<td>Lec 8: Join Algorithms</td>
</tr>
<tr>
<td>W5</td>
<td>Oct. 3</td>
<td>Lec 9: Query Optimization</td>
</tr>
<tr>
<td>W6</td>
<td>Oct. 8</td>
<td>Columbus Day</td>
</tr>
<tr>
<td>W6</td>
<td>Oct. 10</td>
<td>Lec 10: Query Optimization 2 – Final Project Proposals</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Due: Final Project Teams (6.830 Only)</td>
</tr>
<tr>
<td>W7</td>
<td>Oct. 15</td>
<td>Lec 11: Query Execution, In-Memory DB</td>
</tr>
<tr>
<td>W7</td>
<td>Oct. 17</td>
<td>Celebration of Knowledge (Quiz 1)</td>
</tr>
<tr>
<td>W8</td>
<td>Oct. 22</td>
<td>Lec 12: Transactions and Locking</td>
</tr>
<tr>
<td>W8</td>
<td>Oct. 24</td>
<td>Lec 13: Concurrency Control and Snapshot Isolation</td>
</tr>
</tbody>
</table>
# Tentative Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>W9</td>
<td>Oct. 29</td>
<td>Lec 14: Recovery</td>
</tr>
<tr>
<td>W9</td>
<td>Oct. 31</td>
<td>Lec 15: Recovery 2</td>
</tr>
<tr>
<td>W10</td>
<td>Nov. 5</td>
<td>Lec 16: Distributed Databases</td>
</tr>
<tr>
<td>W10</td>
<td>Nov. 7</td>
<td>Lec 17: Distributed Transactions</td>
</tr>
<tr>
<td>W11</td>
<td>Nov. 12</td>
<td>Lec 18: Data Integration</td>
</tr>
<tr>
<td>W11</td>
<td>Nov. 14</td>
<td>Lec 19: Advanced DM Topics</td>
</tr>
<tr>
<td>W12</td>
<td>Nov. 19</td>
<td>Lec 20: Advanced DM Topics</td>
</tr>
<tr>
<td>W12</td>
<td>Nov. 21</td>
<td>Thanksgiving</td>
</tr>
<tr>
<td>W13</td>
<td>Nov. 26</td>
<td>Lec 21: Advanced DM Topics</td>
</tr>
<tr>
<td>W13</td>
<td>Nov. 28</td>
<td>Project/Lab discussions</td>
</tr>
<tr>
<td>W14</td>
<td>Dec. 3</td>
<td>Celebration of Knowledge 2 (Quiz 2)</td>
</tr>
<tr>
<td>W14</td>
<td>Dec. 5</td>
<td>Project/Lab discussions</td>
</tr>
<tr>
<td>W15</td>
<td>Dec. 10</td>
<td>Lec 22: Guest Lecture</td>
</tr>
<tr>
<td>W15</td>
<td>Dec. 12</td>
<td><strong>Final Project</strong>/Poster session</td>
</tr>
</tbody>
</table>
6.814

• 4-5 Problem Sets
• 4 Normal Labs
• 2 Special Labs (team size 2-3)
• 2 Quizzes

• Grading:
  – Problem Sets and Labs: 65% total
  – Exams: 15% each
  – Class Participation: 5%

6.830

• 4 Problem Sets
• 4 Normal Labs
• Final Project (team size 2-3)
• 2 Quizzes

• Grading
  – Problem Sets and Labs: 35% total
  – Exams: 15% each
  – Final Project: 30%
  – Class Participation: 5%
Other Stuff

• **Readings**
  – Please do them
  – Weekly questions -- come to class prepared to answer (not going to collect written answers unless needed)
  – If you do the readings, quizzes will be much easier

• **Notes**
  – I will post all slides, but they are NOT self explanatory
  – **You have to take your own notes**

• **Late days**
  – Each student is allowed 5 "late days", each of which may be used to turn in one problem set or lab one day (24 hour period) later than it is due without penalty.
  – After all five late days are used, assignments will be docked one letter grade for each day they are late.
  – Late days may not be used for the final project/lab 6 or exams.
  – PS must be turned before solutions are posted

• **Collaboration Policy**
  – OK discuss, not OK to share
  – Don’t cheat – we use software to detect duplicates
Tools

- **Piazza**: piazza.com/mit/fall2018/68306814
- **Gradescope**: Entry Code: **9K3BWG**
- GitHub
Clicker Question

• http://clicker.csail.mit.edu/6.814/

How do you take class notes:
(A) I prefer pen and paper
(B) I always use my laptop
(C) I never take notes as I have the best memory
Mobile Phones / Laptops

• No mobile phones

  Exceptions have to be at least as important as:
  (1) your girlfriend/wife is about to give birth,
  (2) you have a life or death family emergency,
  (3) you are about to sell your startup and will invite the whole class for
      (alcohol-free!!!) drinks afterwards
Mobile Phones / Laptops

• No mobile phones
  Exceptions have to be at least as important as:
  (1) your girlfriend/wife is about to give birth,
  (2) you have a life or death family emergency,
  (3) you are about to sell your startup and will invite the whole class for
      (alcohol-free!!!) drinks afterwards

• Laptops
  – Only If you use them for notes taking or clicker questions
Mobile Phones / Laptops

- No mobile phones
- By default, laptop use is prohibited.
  - There are only three exceptions:
    - I’ve explicitly given permission to use laptops for some task (e.g., clicker)
    - You have some documentable reason that requires laptop use. If so please discuss it with me beforehand
    - Emergencies (like with mobile phones)

6.4 Research Literature The following are research papers you can read to learn more.
  - In-class laptop use and its effects on student learning by Carrie B. Fried, Computers & Education, 2007
  - Daydreaming and its correlates in an educational environment by Sophie Lindquist and John McLean, Learning and Individual Differences, 2011
  - The impact of laptop-free zones on student performance and attitudes in large lectures by Nancy Aguilar-Roca, Adrienne Williams, and Diane O’Dowd, Computers & Education, 2012
  - Laptop multitasking hinders classroom learning for both users and nearby peers by Faria Sana, Tina Weston, Nicholas J. Cepeda, Computers & Education, 2013
  - The pen is mightier than the keyboard: Advantages of longhand over laptop note taking by Pam A. Mueller and Daniel M. Oppenheimer, Psychological Science, 2014
  - Logged in and zoned out: How laptop internet use relates to classroom learning by Susan Ravizza, Mitchell Uitvlugt, Kimberly Fenn, Psychological Science, 2017
Participation Grade

- Class participation
- Clicker
- Piazza
What Is This Class About
The Unreasonable Effectiveness of Data

Alon Halevy, Peter Norvig, and Fernando Pereira, Google
“Scientific hindsight shows that Google Flu Trends far overstated this year's flu season....”

“Lots of media attention to this year's flu season skewed Google's search engine traffic.”

David Wagner, Atlantic Wire, Feb 13 2013
“The intuition behind this ought to be very simple: Mr. Obama is maintaining leads in the polls in Ohio and other states that are sufficient for him to win 270 electoral votes.”

Nate Silver, Oct. 26, 2012
fivethirtyeight.com

“...the argument we’re making is exceedingly simple. Here it is: Obama’s ahead in Ohio.”

Nate Silver, Nov. 2, 2012
fivethirtyeight.com

“The bar set by the competition was invitingly low. Someone could look like a genius simply by doing some fairly basic research into what really has predictive power in a political campaign.”

Nate Silver, Nov. 10, 2012
DailyBeast
The data-driven game behind Obama's campaign

"In the 21st century, the candidate with [the] best data, merged with the best messages dictated by that data, wins."

Andrew Rasiej, Personal Democracy Forum

“...the biggest win came from good old SQL on a Vertica data warehouse and from providing access to data to dozens of analytics staffers who could follow their own curiosity and distill and analyze data as they needed.”

Dan Woods
Jan 13 2013, CITO Research

“The decision was made to have Hadoop do the aggregate generations and anything not real-time, but then have Vertica to answer sort of ‘speed-of-thought’ queries about all the data.”

Josh Hendler, CTO of H & K Strategies
“Non-Analytics”? Applications
Simple Truths

• „Power of data“
  – the more data the merrier (GB -> TB)
  – data comes from everywhere in all shapes
  – value of data often discovered later
  – data has no owner within an organization (no silos!)

• Services turn data into $$
  – the more services the merrier
  – need to adapt quickly

• E.g.: Google, Amadeus, Disney, Walmart, BMW, ...
• Platforms: Oracle, MS, SAP, Google, ...
Big Data Question: Yes or No?

• Find a spouse?
• 1 + 1?
• Cure for cancer?
• How to treat a cough?
• Should I give Tim a loan?
• Premium for fire insurance?
• When should my son come home?
• Which book should I read next?
• Translate from German to English.
Why Databases?
An Example: Zoo App

• Admin interface
  – Edit / Add animal

• Public
  – Search for animals
  – Maps

• Zookeeper
  – Feeding calendar

1M animals, 5k pages, 10 admins, 200 keepers
Our Zoo

Sam the Salamander
Endangered species
cage 10

Cheng the Giraffe
Extra tall cage 2

Sally the Student
Classroom Cage 155

Zookeepers
Tim
Raul

Raul feed knowledge every Monday at 2:30pm to Sally

Tim feed “best 6.830 Stories” to Sam every day at 1pm
How Would You Build the Application Without a Data Management System
What is a Database Management System (DBMS)?

• A DBMS is a tool that helps develop and run data-intensive applications:
  – large databases
  – large data streams
Overview of Data Models

• Network model (e.g., CODASYL/COBOL)
• Hierarchical model (IBM IMS/FastPath)
• **Relational model** (SQL)
• Object-oriented model (ODMG 2.0)
• Semi-structured model (XML Infoset)
• Deductive model (Datalog, Prolog)
Animals have names, ages, species
Keepers have names
Cages have cleaning times, buildings
Animals are in 1 cage; cages have multiple animals
Keepers keep multiple cages, cages kept by multiple keepers
Queries: Names of Giraffes

• Imperative
  for each row r in animals
    if r.species = 'giraffe'
    output r.name

• Declarative
  SELECT r.name FROM animals
  WHERE r.species = 'giraffe'
Average Age of Bears

• Declarative

```sql
SELECT AVG(age) FROM animals
WHERE species = 'bear'
```
Cages in Building 32

• Imperative
  
  for each row a in animals
  for each row c in cages
  if a.cageno = c.no and c.bldg = 32
  output a

• Declarative
  
  SELECT a.name FROM animals AS a, cages AS c
  WHERE a.cageno = c.no AND c.bldg = 32
Why use a DBMS?

• Avoid redundancy and inconsistency
• Rich (declarative) access to the data
• Synchronize concurrent data access
• Recovery after system failures
• Security and privacy

• Reduce cost and pain to do something useful
  – There is always an alternative!!!
Fundamental Concepts

• **Data modeling & layout**
  – Systematic approach to structuring / representing data

• **Declarative Querying and Query Processing**
  – High level language for accessing data:
    “Say what I want, not how to do it”
  – "Data Independence"
  – Compiler that finds optimal plan for data access
  – Many low-level techniques for efficiently getting at data

• **Consistency / Transactions + Concurrency Control**
  – **Atomicity** -- Complex operations can be thought of as a single atomic operation that either completes or fails; partial state is never exposed
  – **Consistency and Isolation** -- Semantics of concurrent operations are well defined -- equivalent to a serial execution, respecting invariants over time
  – **Durability** -- Completed operations persist after a failure
Database Abstraction Layers

Data Independence

Logical Data Independence

Physical Data Independence

Physical Layer (e.g., indexes)

Logical Layer (schema)

View1 → View 2 → ... → View 3

Changes at one layer do not affect another layer!
Components of a Database System

- Application
- Ad-hoc Query
- Compiler
- Management tools
- DML-Compiler
- DDL-Compiler
- Query Optimizer
- Runtime
- Schema
- DBMS
- Storage Manager
  - Logs
  - Indexes
  - DB
  - Catalogue

User Types:
- "Naive" User
- Expert User
- App-Developer
- DB-admin

TA Management
Recovery

Application Ad-hoc Query Compiler Management tools DML-Compiler DDL-Compiler Query Optimizer Runtime DBMS Schema Storage Manager Logs Indexes DB Catalogue External Storage
Typical Applications
(data / operations)

- Bank (Accounts / „Money Transfer“)
- Library (Books / „Lend Book“)
- Content Management System (docs, „show“)
- E-Business (Catalogue, „search“)
- ERP (Order, „delivery“)
- Decision Support (Order, „emp of the month“)
- Facebook, Twitter, ... (Friends, „post tweet“)
Data Science: Science of Questions

• How to formulate questions?
  – relational algebra

• How to organize data to answer questions?
  – ER / UML, relational data model

• How to acquire data to answer questions?
  – project, transactions, (much more not covered)

• How to make it efficient
  – normal forms, optimization

• How to quantify error, avoid stupid questions?
  – not covered in this class 😞