Welcome to CPSC 404 Advanced Relational Databases

Instructor: Laks V.S. Lakshmanan

Email: laks@cs.ubc.ca

Office: ICICS/CICSR 315-2366 Main Mall

Lectures: M,W,F: 9-9:50 am, DMP 110.

Office Hour: **see**
http://www.cs.ubc.ca/~laks/404.html.

TAs: Mohammed Alam & Min Xie
(malam@cs.ubc.ca & minxie@cs.ubc.ca).

Why care about DB technology? 1/3

- One of the most successful industries.
- What powers your ATMs, or e-commerce portals, or web services, …?
- What happened with Royal Bank's infamous "software glitch" in June 2004?
  - Customer transactions, incl. payroll deposits not reflected in account balances over several days.
  - Fraudsters trying to cash in on the opportunity.
  - Spillover effect on BMO and TD customers!

Why care about DB technology? 2/3

- **Social Networking & Recommender Systems**: DBMS – Underlying core powering facebook, myspace, flickr, del.icio.us, Yahoo!Answers, rottentomatoes.com, ….
- Pretty much any interesting application of computing, at its core, represents and manipulates data.
- data management will remain important for ever:
  - Continued improvement/extensions of relational technology.
  - Developing technologies for managing data not managed (well): e.g., text, multimedia, web data, graphs, matrices, …

Why care about DB technology? 3/3

- **"Data is the Next Intel Inside**
  - Every significant internet application to date has been backed by a specialized database: Google's web crawl, Yahoo!'s directory (and web crawl), Amazon's database of products, eBay's database of products and sellers, MapQuest's map databases, Napster's distributed song database. As Hal Varian remarked in a personal conversation last year, "SQL is the new HTML." Database management is a core competency of Web 2.0 companies, so much so that we have sometimes referred to these applications as "infoware" rather than merely software. ..." -- **What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software** *(Tim O. Reilly)* .

Course Material

- Text*: R. Ramakrishnan and J. Gehrke, *Database Management Systems*, McGraw-Hill, 3rd Ed., 2003. (preferred).
- What if you have already bought the 2nd edition?
  - Don't despair! You can make do with it. (May need to consult 3rd edition from time to time.)
  - Table of correspondences coming up.
- References:
  - R1: H. Garcia-Molina, J.D. Ullman, and J. Widom, *Database System Implementation*, Prentice Hall, 2000.

    OR

    R2: H. Garcia-Molina, J.D. Ullman, and J. Widom, *Database Systems, The Complete Book*, Prentice Hall, 2002.
  - R3: H. Korth, A. Silberschatz, and S. Sudarshan, Database System Concepts, McGraw-Hill, 6th Ed., 2010.

Both Text and R2, R3 will be available on course reserve from ICICS Reading Room.

Course Material

- R4: For Locality Sensitive Hashing: Ch. 3 of Anand Rajaraman and Jeffrey D. Ullman. Mining Massive Data Sets. http://i.stanford.edu/~ullman/mmds.html

Course Material -- Objectives

- 304 is about basic relational DB design, DB use, and programming
- 404 is meant to "open the black box"
  - Particularly how to tune the performance of the DBMS
  - E.g., what to do if DB requirements/workload change? What index to create? etc.
  - For DBA (vs database programmer)
  - Newer applications (time permitting).

Topics 1/2

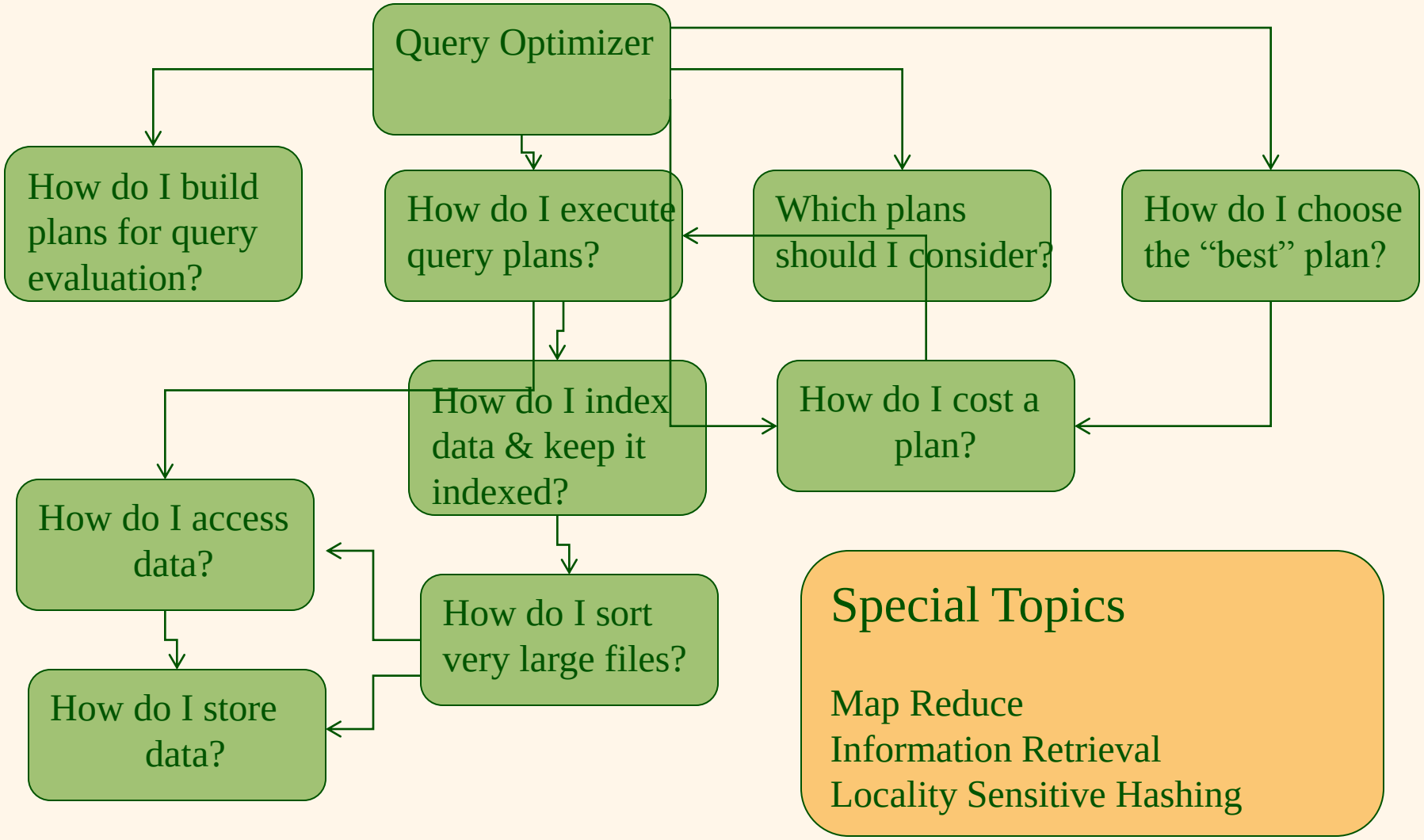
No.	Topic	Text (3rd Edn.) Chapter(s)	2nd Edn. * Chapter(s)
1.	Review	9	7
2.	External Sorting#	13	11
3.	Tree-structured Indexing	10	9
4.	Hash-based Indexing	11	10
5.	Query Evaluation & Optimization	12	13
6.	Q E & O	14	12
7.	Q E & O	15	14
8.	Map Reduce	--	--
9.	Info. Retrieval	27	22
10.	Locality Sensitive Hashing	--	--

*Coverage may be inadequate. #Reading Assignment. ·

Topics 2/2

- External Sorting: draw upon R2, Ch: 11.4.
  - This time, ES will be mainly assigned reading, for self study, with an overview and summary from me.
- If at all you are using the 2nd edition of text (discouraged), be sure to consult the 3rd edition from time to time.
- For Locality Sensitive Hashing, we will draw upon Ch. 3 of R4.

How do they tie together?



Am I prepared for CPSC 404?

- CPSC 304 background assumed in an essential way.
- No time to review 304 in class: course will be relatively fast paced.
- But, you **must** refresh 304 material and be prepared to answer questions based on 304.
- Take the time to read course outline (these slides) carefully.
- Make sure you understand assumptions and obligations. (Ask any questions you may have, early!)
- Make sure you are aware of resources available for help.

About Lectures, Notes, etc.

- Lectures need **not** follow text closely, although materials are compatible
- Notations may differ
- You are responsible for the text, appropriate reference chapters, lectures, and any additional reading that may be assigned
- Lecture notes available at http://www.cs.ubc.ca/~laks/404notes.html
- Parts of some slides may be blank (in the notes). This is intentional: the blanks will be filled (only) in class. If you miss the class, get the material from a friend: the online notes will NOT contain the filled material.
- Some material presented in class (e.g., on write-on transparencies or on board) may NOT appear in the online notes.

What resources are available for help?

- Course home page: http://www.cs.ubc.ca/~laks/404.html, Visit it often for important announcements/info.
- Make sure your email address registered with SSC is valid and working.
- Online notes: http://www.cs.ubc.ca/~laks/404notes.html
- My office hours: group mode as needed.
- TA: office hours/email; see course home page for details.
- NOTE: We will use **piazza for all online 404-related discussion**. TAs and I will be monitoring piazza. For questions related to the course, piazza is your best bet to get the answers as soon as possible.
- Email minxie@cs.ubc.ca to join the piazza group for 404.

About assignments, quizzes, final 1/3

- Assignments:
  - Watch for assignment box details on course home page.
  - due NO LATER THAN 5 pm on the due date.
  - Late submissions levied a penalty of 10%/day.
  - Not accepted after 3 days past due date.

About assignments, quizzes, final 2/3

- Quizzes:
  - coverage typically incremental and up to last lecture of previous week.
  - We may require assigned seating (watch for announcements).
  - We will require you to sign an honor code.
  - Absence must be explained with proper documentation:
    - E.g., doctor's note for health related absence.

About assignments, quizzes, final 3/3

- Final typically will cover whole course.
- Please do not leave room after quiz/final until you are instructed to, even if you have finished and handed in your exam.

About Cheating

- Cheating is a serious offence at UBC. Be aware of its seriousness and the penalty it will attract:
  - E.g., copy or plagiarize parts of an assignment from another student ➔ zero course mark & suspension for 4 months
  - E.g., cheat in midterm ➔ zero course mark & suspension for 8-12 months
- See "Student Discipline Report", *Sept. 2005-Aug. 2006.* www.universitycounsel.ubc.ca/discipline/05-06.pdf & http://www.cs.ubc.ca/about/policies/collaboration.shtml
- Remember: You are responsible for knowing what constitutes cheating. And cheating stinks!
- Take a look at the following document written by Prof Tamara Munzner: http://www.cs.ubc.ca/~tmm/courses/cheat.html

	Week of	Monday	Wednesday	Friday
	Sept. 3	X	Outline/Review1	Review1
	Sept. 10	Sorting	Btree	Btree
Asst #1 →	Sept. 17	Btree	Btree	Btree/Hashing
	Sept. 24	Hashing	Hashing	Hashing
	Oct. 1	Hashing	QE	QE
Asst #2 →	Oct. 8	X	QE	QE
	Oct. 15	Quiz1	Optimize	Optimize
	Oct. 22	Optimize	Optimize	Optimize
	Oct. 29	Map Reduce	Map Reduce	Map Reduce
Asst #3 →	Nov. 5	Map Reduce	Map Reduce	Quiz2
	Nov. 12	X	IR	IR
	Nov. 19	IR	IR	LSH
	Nov. 26	LSH	LSH	LSH
	Dec. 3	Review (tentative)	X	X

Tentative Schedule

Course Evaluation

	Percentage
Final exam	45%
2 in-class quizzes	40%
3 assignments	15%

- In addition, in-class problem solving (participation required):
  - Call at random
  - In several groups of 2-3 (neighbors)
  - One randomly chosen solution will be discussed
  - Solvers' identity anonymous
- Why bother?
  - Everybody learns; sometimes more from mistakes

Course Notes

- All notes on the web: http://www.cs.ubc.ca/~laks/404notes.html
- Extra in-class examples (which will not be in the online notes).
- Blanks in notes will only be filled in class and will not be reflected in online version.
- Any questions about course policy? Raise policy questions early.

Beyond CPSC 404 – Extra Credit

- Why: encourage motivated students to go beyond classroom and course
- Who: those interested in higher studies or just interested in knowing about cutting edge topics in data management & data mining.
- What: read papers on special topics, discuss, and critique. Possibly work on specific research problems with me.
- Attend "db talks" and "social networking" reading groups (Time TBD); possibly make presentations.
- No course marks for this exercise
- Will reflect in reference letters, though
- And if you are up for it, you will get much more value.
- Interested? Email me (laks@cs).