Introduction to Data Management
CSE 344

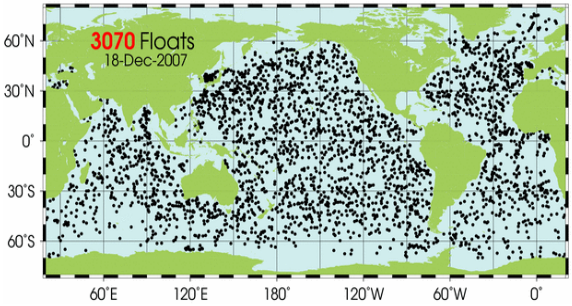
Lecture 1: Introduction

Staff

- Instructor: Sudeepa Roy
  - sudeepa@cs.washington.edu
  - Office hours: Wednesdays, 3:30-4:20, in CSE 344 (my office) ☺

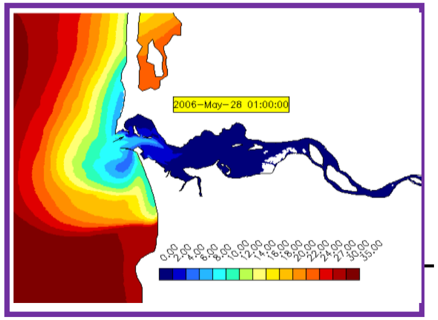
- TAs:
  - Aloka Krishnan, alokak@uw.edu,
    Office hours: Tuesday: 2:00-2:50 and Friday 12:20-1:20, CSE 218
  - Vaspol Ruamviboonsuk, vaspol@cs
    Office hours: Monday, 10:30 - 11:20, CSE 218
  - Yi-Shu Wei, yishuwei@cs
    Office hours: Tuesday 10:00-10:50 and Thursday 2:30-3:20, CSE 218

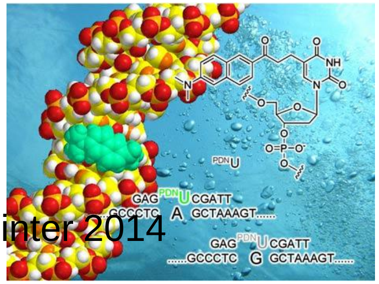
Class Goals

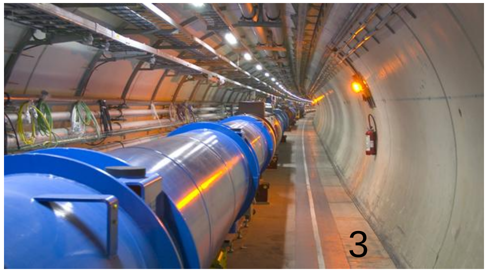


- The world is drowning in “big data”!
  - Web (Google, Facebook, Twitter, News articles), smart devices, sensors, scientic experiments, satellites, …
- Need computer scientists to help manage this data
  - Help domain scientists achieve new discoveries
  - Help companies provide better services (e.g. Facebook)
  - Help governments become more efficient

Large Synoptic Survey Telescope
LSST Primary/Tertiary Mirror Blank







Class Goals

- Welcome to 344: Introduction to Data Management!
  - Existing tools PLUS data management principles

- Next steps:
  - CSE 444: build data management systems
  - CSE 446: learn interesting facts from data

Course Format

- Lectures MWF, 2:30pm-3:20pm, MLR 301
- Sections:
  - AA: Th 12:30-1:20 EEB 037
  - AB: Th 1:30-2:20 EEB 026
  - Content: exercises, tutorials, questions

- 8 Homework assignments
- 7 Web quizzes
- Midterm and Final

Communications

- Web page: http://www.cs.washington.edu/344
  - Syllabus is there
  - Lectures will be available there (see calendar)
  - Homework assignments will be available there
  - Link to web quizzes is there
- Mailing list
  - Announcements, group discussions
  - You are already subscribed

Communications

- Discussion board
  - Great place to ask assignment-related questions…
  - …also to discuss concepts
  - Post questions here in respective discussion areas for fastest response instead of emails
  - But never post partial/full solutions!

Textbook

Main textbook, available at the bookstore:

- *Database Systems: The Complete Book*,
  Hector Garcia-Molina,
  Jeffrey Ullman,
  Jennifer Widom

  Second edition.

Most important: COME TO CLASS ! ASK QUESTIONS !

Other Texts

Available at the Engineering Library

(not on reserve):

- *Database Management Systems*, Ramakrishnan
- *XQuery from the Experts*, Katz, Ed.
- *Fundamentals of Database Systems*, Elmasri, Navathe
- *Foundations of Databases*, Abiteboul, Hull, Vianu
- *Data on the Web,* Abiteboul, Buneman, Suciu

Grading

• Homeworks	30%
• Web quizzes	20%
• Midterm	20%
• Final	30%

Eight Homeworks

H1&H2: Basic SQL with SQLite

H3: Advanced SQL with SQL Server

H4: Relational algebra, Datalog

H5: XML and XQuery with Saxon

H6: Conceptual Design

H7: SQL in Java (JDBC)

H8: Parallel processing with MapReduce

Homeworks (except HW2) are due Thursday night – dropbox!

About the Homeworks

- Homework assignments will take time but most time should be spent *learning*

- Must be done on your own!

- Very practical assignments

- Put everything on your resume!!!
  - SQL, SQLite, SQL Server, SQL Azure, JDBC, XML, XQuery, Saxon, Amazon Elastic MapReduce, Hadoop, Pig Latin, …

Late Days

Max 4 late days per quarter;
Max 2 per homework

- in 24 hours chunk
- submission between 12:00 am to 11:59 pm next day counts as one late day

Late days = safety net, not convenience!

- Normally, you should use zero late days
- If you have an emergency during the quarter, you should use 1 or 2.
- If you use all 4, you are doing it wrong.
- **No late day for HW8!**

Seven Web Quizzes

- Write down class token (also on discussion board)
- Short online tests
- Can take many times: best score counts!
- **No late days!** Gradiance will close at 11:59 pm on due dates
- But lowest score is dropped!
- Provides explanations for wrong answers
- Will help you
  - Test your knowledge
  - Stay in synch with class
  - Get ready for homeworks

Due Tuesday night (except WQ7)

Exams

- Midterm (02/19) and Final (03/18)
- Open book, open notes (no computers!)
- Check course website for dates
- Location: in class

To Conclude Logistics..

- Attend all lectures and sections
- Ask questions in class
- Come to office hours (at least one everyday!)
- Give us feedback

Outline of Today's Lecture

- Overview of database management systems
  - Why they are helpful
  - What are some of their key features
  - What are some of their key concepts

Database

What is a database ?

Give examples of databases

Database

What is a database ?

- A collection of files storing related data

Give examples of databases

- Accounts database; payroll database; UW's students database; Amazon's products database; airline reservation database

Database Management System

What is a DBMS ?

Give examples of DBMSs

Database Management System

What is a DBMS ?

- *A big program written by someone else that allows us to manage efficiently a large database and allows it to persist over long periods of time*

Give examples of DBMSs

  - Oracle, IBM (DB2, Informix), Microsoft (SQL Server, Access)
  - Sybase
  - Open source: MySQL (Sun/Oracle), PostgreSQL
  - Open source library: SQLite

We will focus on relational DBMSs most quarter

An Example: Online Bookseller

- What data do we need?

  –

  –

  –

- What capabilities on the data do we need?

  –

  –

  –

An Example: Online Bookseller

- What data do we need?
  - Data about books, customers, pending orders, order histories, trends, preferences, etc.
  - Data about sessions (clicks, pages, searches)
  - Note: data must be persistent! Outlive application
- What capabilities on the data do we need?
  -
  -
  -

An Example: Online Bookseller

- What data do we need?
  - Data about books, customers, pending orders, order histories, trends, preferences, etc.
  - Data about sessions (clicks, pages, searches)
  - Note: data must be persistent! Outlive application
- What capabilities on the data do we need?
  - Insert/remove books, find books by author/title/category/price, create order history, sales
  - Find popular books; recommend books
  - Note: data must be accessed efficiently, by many users

Multi-user discussion

- Jane and John both have ID number for gift certificate (credit) of $200 they got as a wedding gift
  - Jane @ her office orders "The Selfish Gene, R. Dawkins" ($80)
  - John @ his office orders "Guns and Steel, J. Diamond" ($100)

- Questions:
  - What is the ending credit?
  - What if second book costs $130?
  - What if system crashes?

DBMS Benefits

- Expensive to implement all these features inside the application
- DBMS provides these features (and more)
- DBMS simplifies application development

Client/Server Architecture

- One *server* that stores the database (DBMS):
  - Usually a beefy system
  - But can be your own desktop…
  - … or a huge cluster running a parallel DBMS
- Many *clients* run apps and connect to DBMS
  - E.g. Microsoft's Management Studio
  - Or psql (for PostgreSQL)
  - Or some Java/C++ program
- Clients "talk" to server using JDBC protocol

People

- **DB designer**: establishes schema (344)
- **DB administrator**: loads data, tunes system, keeps whole thing running (344, 444)
- **DBMS implementor**: builds the DBMS (444)
- **DB application developer**: writes programs that query and modify data (344)
- **Data analyst**: data mining, data integration (344, 446)

Key Data Mngmt Concepts

- **Data models**: how to describe real-world data
  - Relational, XML, graph data (RDF)
- **Schema v.s. data**
- **Declarative query language**
  - Say what you want not how to get it
- **Data independence**
  - Physical independence: Can change how data is stored on disk without maintenance to applications
  - Logical independence: can change schema w/o affecting apps
- **Query optimizer** and compiler
- **Transactions**: isolation and atomicity

What This Course Contains

- **Focus: Using DBMSs**
- Relational Data Model
  - SQL, Relational Algebra, Relational Calculus, datalog
- Semistructured Data Model
  - XML, XPath, and XQuery
- Conceptual design
  - E/R diagrams, Views, and Database normalization
- Transactions
- Parallel databases, MapReduce, and Pig-Latin
- Data integration and data cleaning

What to Do Now

http://www.cs.washington.edu/344

- Webquiz 1 is open
  - Create account at http://newgradiance.com/
  - Use course token
  - Webquiz due next Tuesday
- Homework 1 will be posted tomorrow
  - Simple queries in SQL Lite
  - Homework due next Thursday