Scalable Web Software
CS193S - Jan Jannink - 1/07/10
Administrative Stuff

- Computer Forum Career Fair: Wed. 13, 11-4
- Lawn between Hewlett Teaching Center and Gilbert Building
- Looking forward to your emails! Already received a dozen or so
- Any problems dealing with eclipse/gwt setup?
1. Scalability: (Jan.)
2. Agile Practices
3. Ecology/Mashups*
4. Browser/Client
5. Data/Server: (Feb.)
6. Security/Privacy
7. Analytics*
8. Cloud/Map-Reduce
9. Publish APIs: (Mar.)*
10. Future

* assignment due
Quick Review

• Think Big

• Scalability means thriving and growing in a dynamic environment

• Java + open source: a great environment to learn scalable practices

• Engineers & investors understand innovation differently
how do you get change for this?
The Web is Just Right

- Many basic computer science algorithms are well understood
- Complexity theory defines practically unsolvable problems
- Scalable web coding fits in the border region between basic and complex
YouTube Example

- Piggybacked growth on MySpace.com
- Previously unheard of bandwidth
- Growth delayed feature development
- Could not afford to continue to grow on its own
- Acquisition significantly delayed revenue model development
Properties of Web Data

• Quickly, iteratively, accumulated by contributors, seen by many viewers
• New additions are not independent of prior ones
• The dependency induces several power law distributions over the data
• Frequently called ‘Long Tail data’
Power Law Distributions

• Every aggregated collection of human artifacts seems expressible in this way
• They are so called because in a log/log plot they generally follow a straight line
• the slope of the line $s$ corresponds to a power coefficient: i.e., $y = x^{-s}$
Web Traffic Referral Plot

Number of Referred Users

Referring Page (Rank-Ordered)
Why This is Relevant

- Back end infrastructure
- Database partitioning
- Data replication
- Caching technique
- Web page design
Other Examples

- Human Languages (word usage)
- Social Networks (connections)
- Financial Markets (transactions)
- Research Citations
- Called them Nexus in my PhD research
Power Law Generators

- item creation
- object in a universe
- link creation
- any relationship between items
- new link probability based on existing structure
- item & link removal
Stuff to Ponder
Dynamic Environments

- Habitats/Ecosystems
  - species surviving to occupy a niche
- Organizations/Societies
  - unifying force and leadership
- Markets/Economies
  - producer consumer relationship
• So many technologies, so little time
• Start with a single platform, language
  • Java/gwt/junit
• Single developer code/test cycle
  • code/checkin/test/repeat
• We’ll continue today with SQL
MySQL Install

- download from website & install, download & unzip popnames.zip
- minor Mac OS X command line edit
  - `sudo echo /usr/local/mysql/bin > /etc/paths.d/mysql`
  - `mysql < popnames.txt`
    - `select name, count(name), sum(count) from popnames group by name order by count(name) desc limit 20;`
    - `select gender, name, sum(count) as total from popnames where year(year)>1958 group by name, gender order by total desc limit 20;`
    - `select left(name,1) as i, sum(count) from popnames group by i;`
Untaught MySQL Tools

- `mysqlimport`
- `mysqldump`
- `workbench`
- `connectors`
- `in mysql client`
  - `data analysis : group by, order by`
Popular Names Dataset

- Source data:

- Command line tools extract data from html tables, into tab separated text files
  - `curl`, `sed`, `awk`, `python`

- Import into MySQL
  - `load data local infile "" into table ... fields terminated by "" (@A, @B, @C) set ...;`
User Account Table

- name
- hashed password
- identifier (email)
- current email
- payment info
- (secure stuff)

- everything else (potentially public data) should be kept separate
Q & A Topics

- Nexus research
- Database vs. flat file
- Hibernate vs. ODBC access layer
Worth Checking Out

• MySQL downloads
  • http://dev.mysql.com/downloads/

• The Goldilocks Enigma
  • Paul Davies

• SQLite
  • http://www.sqlite.org/