The web is replacing the desktop

SDKs & toolkits are there

WASHINGTON TECHNOLOGY industry association

What about the backend?

Image: Wikipedia user Calyponte

Two key concepts

- Processing 1000x more data doesn't have to be 1000x harder
- Cycles and bytes, not hardware, are the new commodity

Cloud computing is:

(Engineering definition)

Providing services on virtual machines allocated on top of a large physical machine pool

Cloud computing is:

(Business definition)

A method to address scalability and availability concerns for large scale applications

Cloud computing is:

(The big picture)

Democratized distributed computing

Outline

- Introduction
- Large-Scale Data Processing
- Cluster Management
- Conclusions

Outline

- Introduction
- Large-Scale Data Processing
  - MapReduce
  - The Google File System
  - BigTable
  - Hadoop
- Cluster Management
- Conclusions

What could you do…

…with 1000x as much data & CPU power?

- Scale your business to 1000x more users
- Gather statistics about every user click
- Improve recommendation systems
- Model better price plan choices
- Simulate 1,000,000 users of a system

Why can't you do it?

- Large-scale cluster reliability is hard!
  - Machines fail
  - Hard drives crash
  - Network goes down
  - Software bugs
  - Gremlins?
- Two tasks for two types of people
  - Designing reliable, scalable system architecture
  - Writing data processing algorithms

An old problem…

Image: Bill Bertram

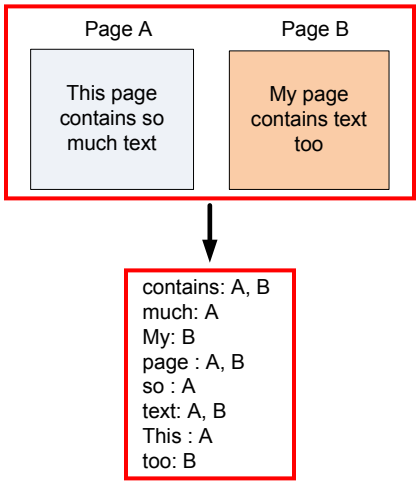
…At a larger scale

Image © Google, Inc.

MapReduce

- Developed at Google in 2003
- Divides *performing computation reliably* from *computing over large data*
- Provides reliable, scalable platform on which to develop distributed applications
- Simplifies programming model for distributed application writers
  - Data flow programming

Example use: Make an index

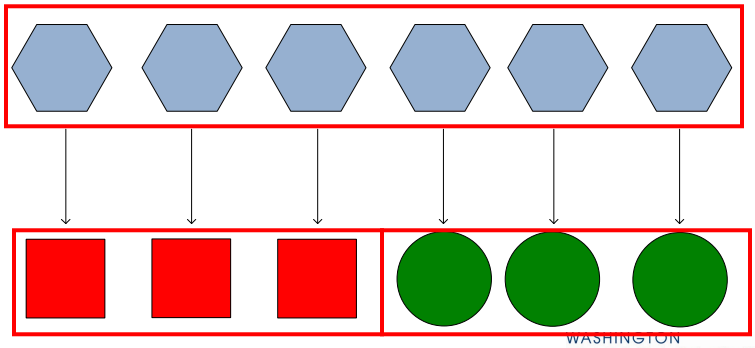


Algorithm:

1. For each page, get a list of all words on the page
2. For each word from (1), get list of pages on which it appears

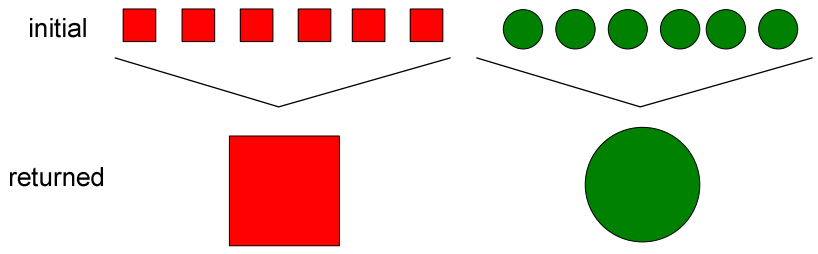
Map

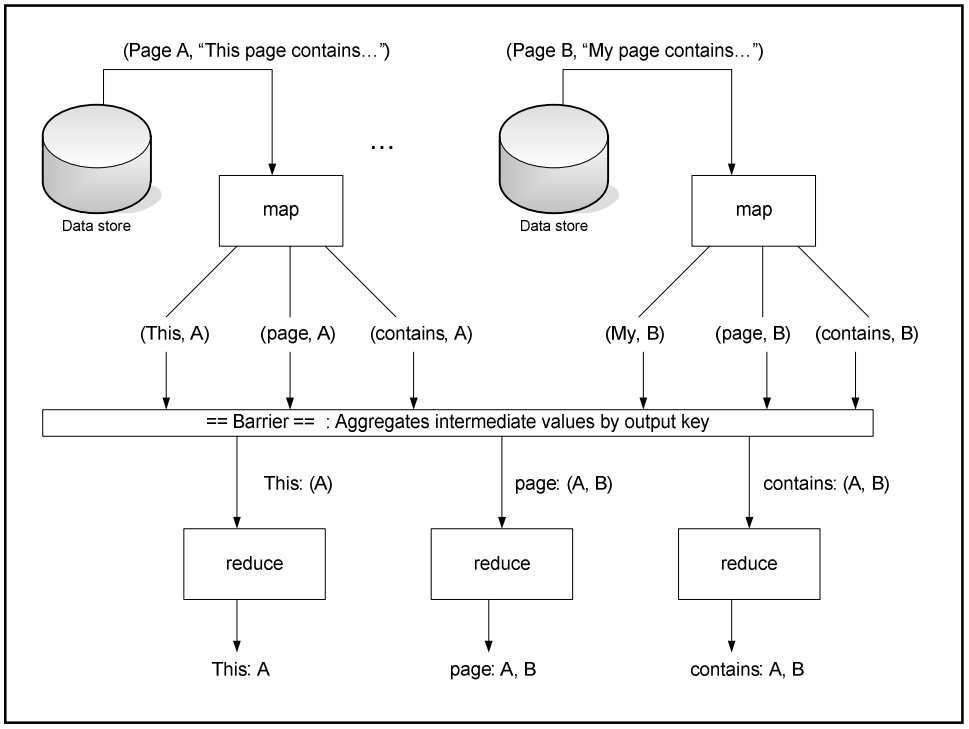
```
map  (in_key, in_value) ->
   (out_key, intermediate_value)
```



Reduce

```
reduce (out_key, intermediate_value list) ->
     out_value
```





Building the index

Page A

This page contains so much text

Page B

My page contains text too

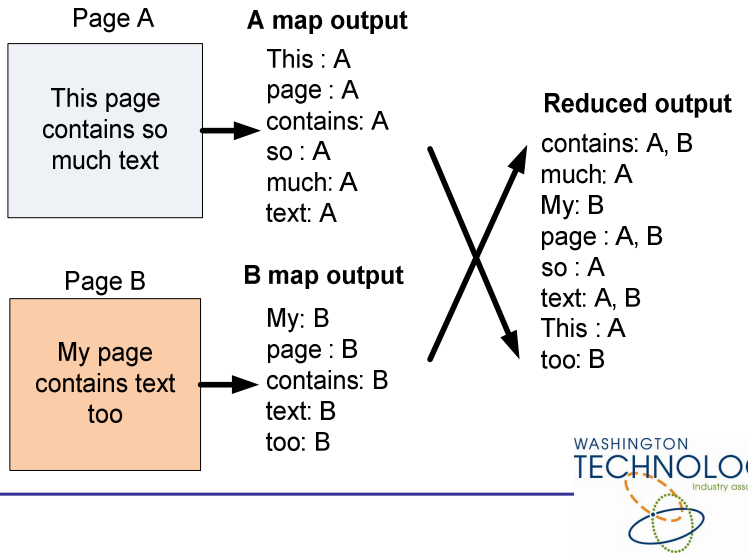
contains: A, B
much: A
My: B
page : A, B
so : A
text: A, B
This : A
too: B

Index: MapReduce

```
map(pageName, pageText):
    foreach word w in pageText:
      emitIntermediate(w, pageName);
    done

reduce(word, values):
    foreach pageName in values:
      AddToOutputList(pageName);
    done
    emitFinal(FormattedPageListForWord);
```

Index: Data Flow



Other applications

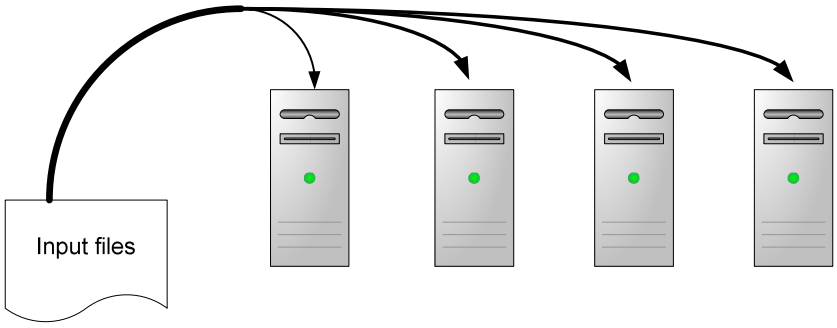
- Classifying instances; machine learning
- Log analysis: click streams, user trends
- Image processing
- A scalable implementation of your backend business logic

The Google File System

The distributed file system is what makes MapReduce work

- Data spread evenly throughout cluster
- Replicated 3x for redundant storage
- Master machine detects failures and rebalances data on the fly

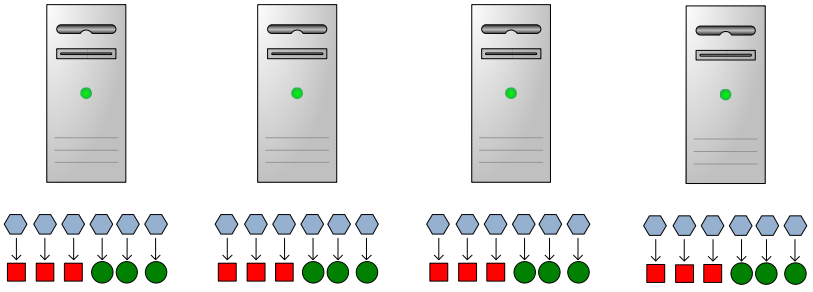
Putting it together: active storage



Data automatically distributed to nodes at load time

WASHINGTON TECHNOLOGY industry association

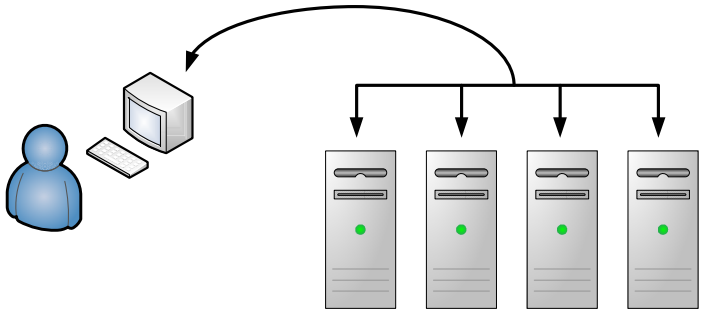
Automatic parallel processing



Data elements processed locally, in parallel

WASHINGTON TECHNOLOGY industry association

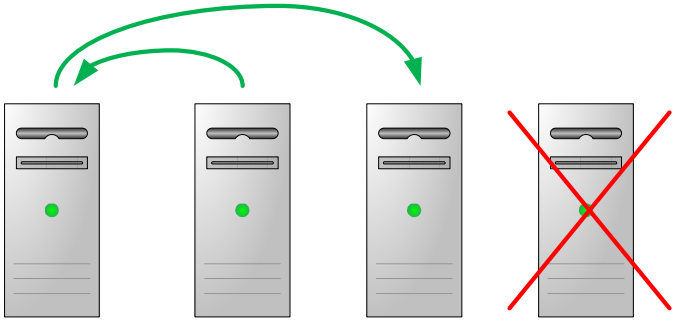
Distributed data, single volume



Output data is written to local disks, and forms a **single user-accessible volume**

WASHINGTON TECHNOLOGY industry association

A self-healing system



Loss of nodes causes automatic data rebalance

WASHINGTON TECHNOLOGY industry association

Adding structure

- GFS provides storage for very large files
  - But structure within files is still slow to use
- Organized database storage still valuable
  - How to get scalable storage and performance?

What does a database do?

- Stores large amounts of data
  - Individual “records” not too large
  - Many records stored per table
- Search over this data (SQL queries)
  - Individual record via index (fast)
  - Complicated joins of related information
- Consistent modifications (transactions)

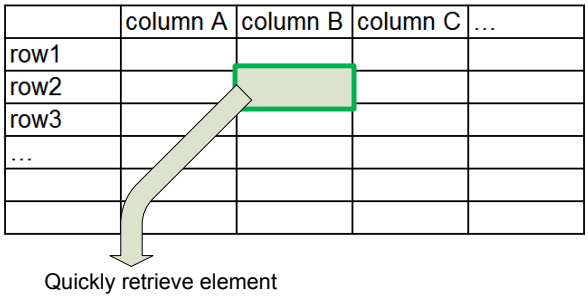
Web scale database problems

- Number of records gets *very* large
- Must be split across several machines
  - Must be stored reliably, redundantly
  - How to keep copies consistent?
- Must have fast access to individual elements, and be able to search it

BigTable

- A Google database layer
- Key idea: divorce organized storage from query system

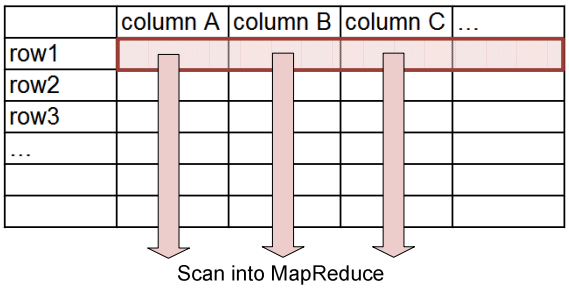
Fast single-element access



- High-speed lookup of individual (row, column)
- Selects data needed by online applications

WASHINGTON TECHNOLOGY industry association

BigTable as a MapReduce input



- Each row is an input record to MapReduce
- MapReduce jobs can sort/search/index/query data in bulk

WASHINGTON TECHNOLOGY industry association

BigTable conclusions

- BigTable provides large-scale storage, MapReduce provides query backend
- Stores immense amount of structured data
  - Uses Google File System for reliability
- Tradeoff: non-standard data access model

Hadoop

- Open source MapReduce + GFS + BigTable
- Java implementation
- Free for commercial use
- Official Apache Software Foundation project

Broad industry adoption

Flexible interoperability

- C++/Python: Pipes/Streaming
- Distributed database: HBase
- Machine learning: Mahout
- Performance tracking: Ganglia
- Development environment: Eclipse
- Hosting: Amazon Web Services

Benefits of Hadoop

- Open source MapReduce, GFS, BigTable
- Clean programming abstractions
  - Allows rapid prototyping of large-scale computation
- Scalable active storage infrastructure
  - Programs for 10 GB of data work for 10 TB
- Reliable storage of dynamic data
  - 3x replication, continuous monitoring

Outline

- Introduction
- Large-Scale Data Processing
- Cluster Management
  - Clouds & Virtualization
  - Amazon Web Services
  - Rightscale
  - AWS + Hadoop
- Conclusions

Constrained Resources

- Large-data applications are I/O bound
  - Hard drive speed is the limiting factor
- Availability of RAM is 2nd level concern
- Raw CPU speed not usually a factor

Typical hardware profile

- "Commodity" server-class machines
  - Think performance/watt or per $$
- 4 processor cores (not fastest available)
- 4-8 GB RAM
- 2x500 GB SATA hard drives

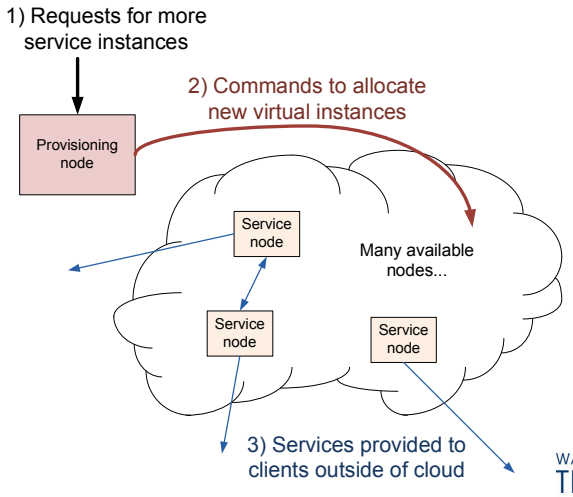
Traditional service architecture

- Expensive upfront, maintenance costs
- Requires many types of resources
  - Machines, power, cooling, bandwidth…
- Does not scale on demand
- Not easily reconfigurable
  - Adding virtualization helps with this

What is a cloud?

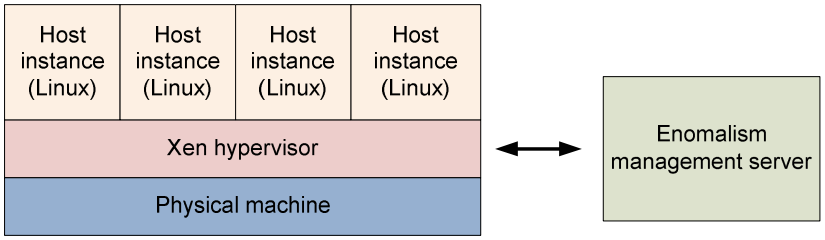
- Virtualized server pool
- Reconfigures to provide different service profiles on demand
- Individual node providing service is unimportant

Clouds: high-level



WASHINGTON TECHNOLOGY industry association

Server architecture



Xen®

enomalism virtualized server management

WASHINGTON TECHNOLOGY industry association

Amazon Web Services

- EC2: Elastic Compute Cloud
- S3: Simple Storage Service
- SQS: Simple Queue Service
- SDB: Simple Database

aws.amazon.com

WASHINGTON TECHNOLOGY industry association

Elastic Compute Cloud

- Provides on-demand processing power
- From $0.10/instance*hour + bandwidth
- Virtual machine images with dynamic or static IP addresses

WASHINGTON TECHNOLOGY industry association

EC2 Templates

- EC2 node configuration stored as *Amazon Machine Image* (AMI)
- 100s of stock AMIs available
  - Generic Linux distributions
  - Hadoop configurations
  - Web server / database installations
- Existing AMIs can be customized & saved for later reuse

Simple Storage Service

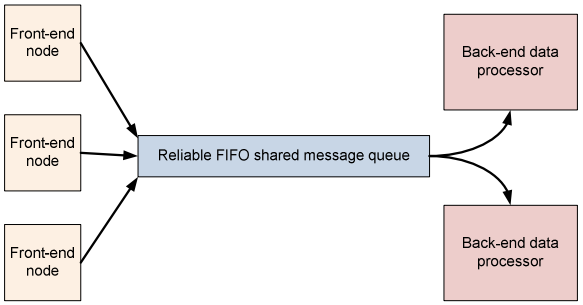
- Virtually infinite storage capacity
- Cost: $0.15/GB*month + bandwidth
  - Free high-speed access to/from EC2 nodes
- Provides permanence layer when EC2 nodes aren't running

Data Security

- Secure data transfer via SSL
- Encrypted storage possible in S3

Simple Queue Service

- Efficient, reliable load distribution layer
- Pay by the message



Simple Database Service

- High availability large-store database
- Provides simple SQL-like language
- Designed for interactive/online use

System reliability

- Geographic server diversity through *availability zones*

Broad existing user base

Powerset
NATURAL LANGUAGE SEARCH

ABACA

37signals

SmugMug

WASHINGTON TECHNOLOGY industry association

Highlight: SmugMug

- Flickr-style online photo hosting
- Entirely stored on S3

They do not own any hard drives!

SmugMug

WASHINGTON TECHNOLOGY industry association

A “web service?”

- Resource provisioning access via SOAP
  - Scriptable client-side tools
  - Not actually exposed on the web by AWS

Web services for the rest of us

RIGHT SCALE

- Web-based interface to AWS
  - Provisioning commands
  - Monitoring & status
- Parameterizable startup scripts

www.rightscale.com

(Rightscale Demo)

RIGHT SCALE

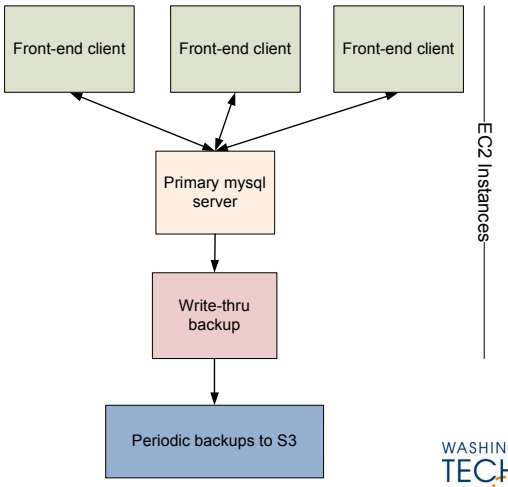
WASHINGTON TECHNOLOGY industry association

Automatic scalability

- Real power of Rightscale is in monitoring
- Scales instances based on demand
- Provides backend reliability in virtual environment

WASHINGTON TECHNOLOGY industry association

mysql setup in EC2/S3



WASHINGTON TECHNOLOGY industry association

Google App Engine

- Integrated cloud computing platform
- Compute resources with GFS- and BigTable-backed storage
- Limited preview now; check back soon!

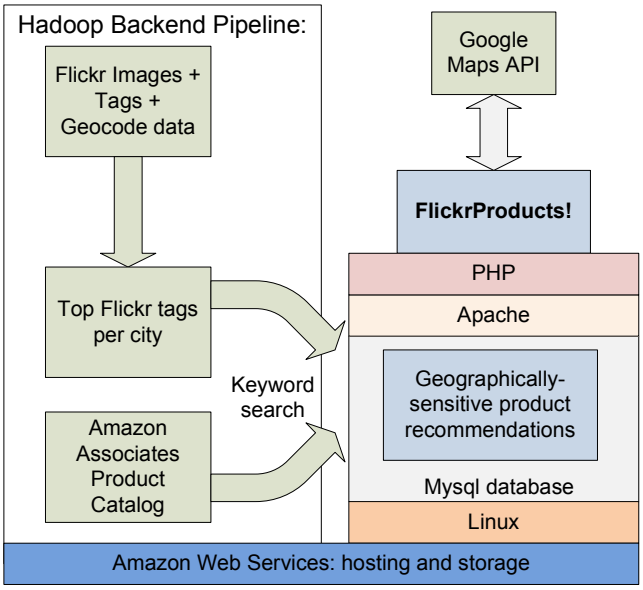
WASHINGTON TECHNOLOGY industry association

Using AWS with Hadoop

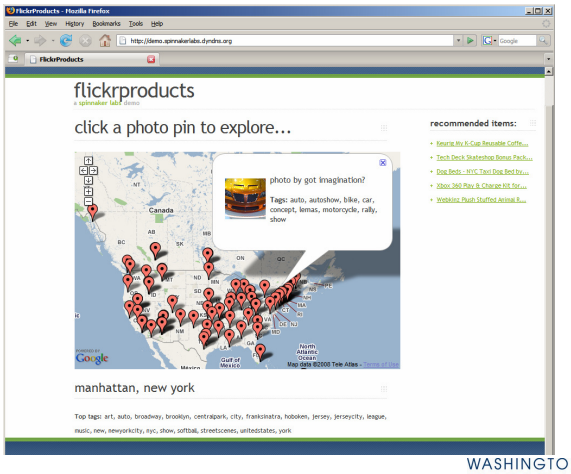
- Hadoop runs on EC2 nodes
  - Templates defined for several versions
- Permanent results storable in S3
- EC2-served frontend retrieves results
  - Published via static IP address

WASHINGTON TECHNOLOGY industry association

Putting it together: FlickrProducts

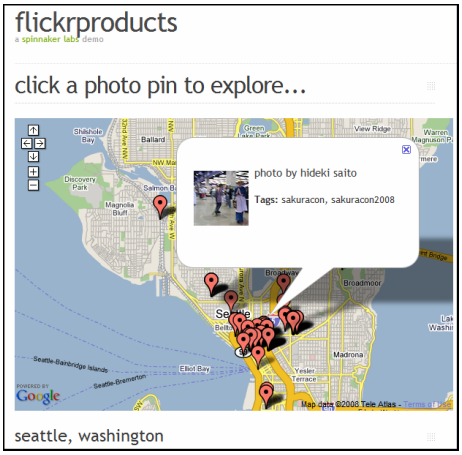


FlickrProducts Tour



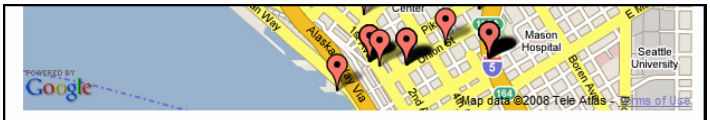
WASHINGTON TECHNOLOGY industry association

Images in context on the map



WASHINGTON TECHNOLOGY industry association

Top tags listed per city

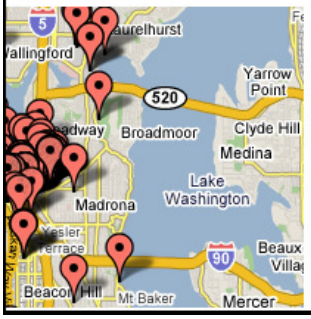


seattle, washington

Top tags: anime, barca, con, convention, cosplay, eyefi, geotagged, kerriesbirthday, march, ontheboards, party, ratcityrollergirls, rcrg, rollerderby, rollergirls, rollerskating, s4b1, sakuracon, sakuracon2008, unitedstates

WASHINGTON TECHNOLOGY industry association

Recommendations from Amazon



recommended items:

+ Hello Kitty Jewelry Box
+ Full Metal Panic! DVD
+ Fruits Basket Cosplay Cap
+ Japanese "Domo-kun" T-shirt

Add to Shopping Cart

WASHINGTON TECHNOLOGY industry association

Demo technical points

- Public Amazon, Flickr APIs for input data
- Hadoop pipeline entirely in Java
- Frontend: mysql + PHP + Javascript
- Full system hosted on AWS EC2
  - VM Image and database backed by S3
- Rightscale management dashboard
- **Conception-to-demo time: two days.**

Conclusions

- Powerful new abstractions for large-scale data processing systems
  - Scalable, reliable, available
  - Support rapid development
- Large managed server pools available
  - Low overhead
  - Eliminate management headaches
  - Grow and shrink according to need

Questions?

Aaron Kimball
aaron@spinnakerlabs.com

WASHINGTON TECHNOLOGY industry association