Cloud Computing @Yahoo!

Raghu Ramakrishnan

Yahoo! Fellow

Chief Scientist, Search and Cloud Platforms

(Many slides courtesy of others at Yahoo!)

Cloud Services @Y!: Use Cases

Yahoo! Data Scale

Massive user base and engagement

- 640M+ unique users, 11B page visits/month
- Hundreds of petabytes of storage
- Hundreds of billions of objects
- Hundreds of thousands of requests/sec, 200B events/day, 200 PB/day

Global

- Tens of globally distributed data centers
- Serving each region at low latencies

Challenging Users

- Rapidly extracting value from voluminous data
- Downtime is not an option (outages cost $millions)
- Variable usage patterns

> “Just look at our homepage, for example. Since we began pairing our **content optimization technology** with editorial expertise, we’ve seen **click-through rates in the Today module more than double**. And we’re making additional improvements to this technology that will make the user experience ever more personally relevant.”
>
> **Carol Bartz, Analyst Call, January 27, 2010**

CONTENT OPTIMIZATION FOR PORTALS

Yahoo! Front Page

Product Objective

Prioritize small pool of editorially programmed packages to optimize engagement in real-time

+79% clicks vs. randomly selected

+160% clicks vs. one size fits all

+43% clicks vs. editor selected

Key Features

Package Ranker (CORE)

Ranks packages by expected CTR based on data collected every 5 minutes

Dashboard (CORE)

Provides real-time insights into performance by package, segment, and property

Mix Management (Property)

Ensures editorial voice is maintained and user gets a variety of content

Package rotation (Property)

Tracks which stories a user has seen and rotates them after user has seen them for a certain period of time

Key Performance Indicators

Lifts in quantitative metrics

Editorial Voice Preserved

Content Optimization & Cloud

Online Learning
- Online regression models, time-series models
- Model the temporal dynamics
- Provide fast learning for per-item models

Offline Modeling
- Exploratory data analysis
- Regression, feature selection, collaborative filtering (factorization)
- Seed online models & explore/exploit methods at good initial points
- Reduce the set of candidate items

Near real-time user feedback

Explore/Exploit
- Multi-armed bandits
- Find the best way of collecting real-time user feedback (for new items)

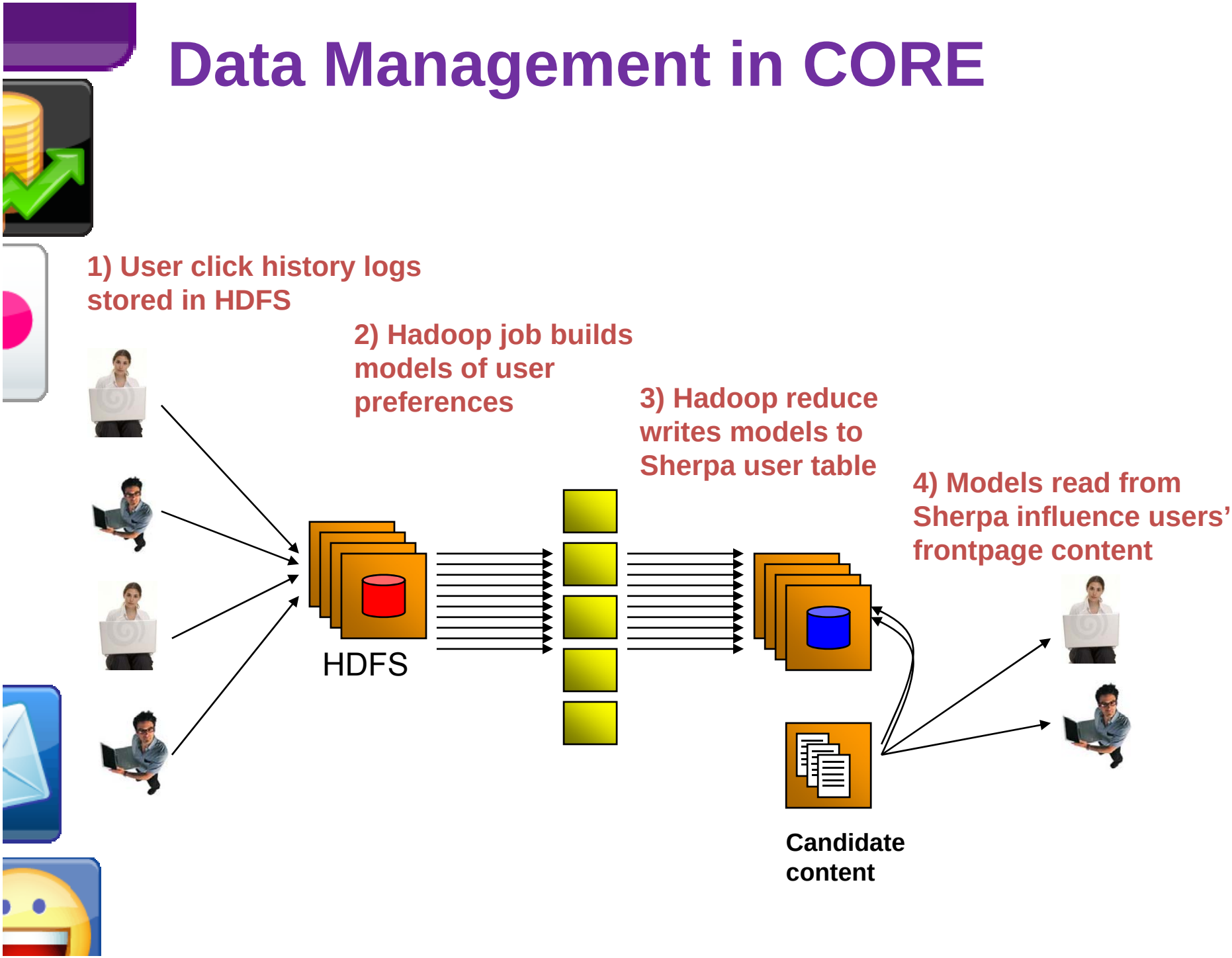
Large amount of historical data (user event streams)

Data pipelines

Hadoop

PNUTS

Data Management in CORE



Matching Users to Content

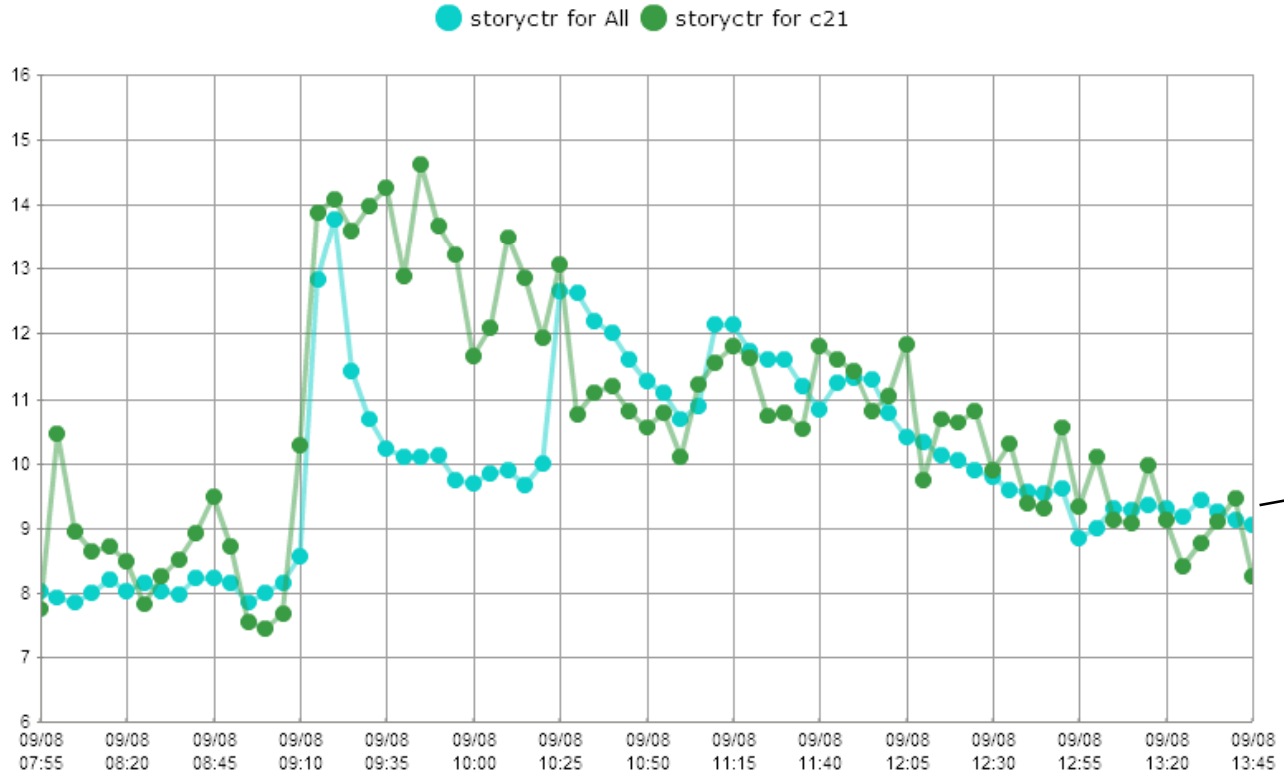
We learn how user attributes correlates with engagement in each item

	Default	Male	Female	18-24	25-34	Heavy Sports
	7.1	-0.4	+0.4	+0.3	+0.1	-0.5
	6.8	+1.0	-1.0	+0.2	+0.3	+2.1
	6.5	-0.6	+0.6	+0.5	+0.3	-0.8
	6.2	0	0	-0.7	-0.5	-0.3
	5.9	-1.1	+1.1	-0.5	-0.2	-0.2
						

We compute rankings for each user based on his/her attributes

CORE Dashboard: Overall CTR

Compare performance of models and historical benchmarks



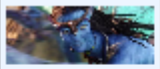
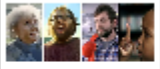
Bucket	Page Views,	All Clicks	Story Clicks	Footer Clicks	OverAll CTR	Story CTR	Footer CTR	Lift compared to All
All	43,005,783	13,061,688	4,286,407	8,775,281	30.37	9.97	20.4	0
c21	281,821	84,853	29,777	55,076	30.11	10.57	19.54	6.02

Compare buckets and models over time

See which content was promoted most across time

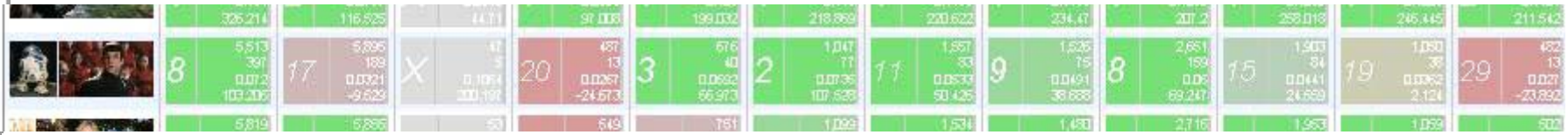
Compare bucket metrics

COKE Dashboard: Segment Heat Map

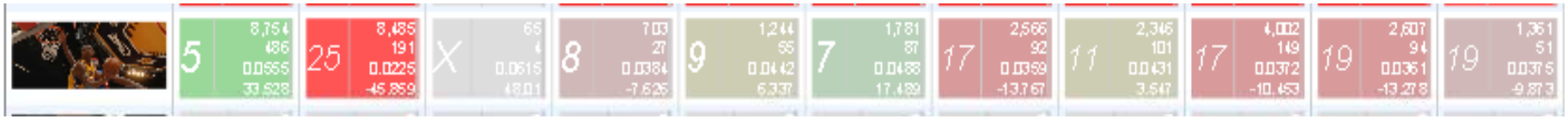
Package	male	female	OMG	BUAuto	BUEnt	BU Fin	Health	BUSport	NBA	BUTrav	ALL
	408,260 18,440 0.0452 8.477	390,404 14,449 0.037 -11.113	270,039 16,940 0.0627 50.661	121,080 7,389 0.061 46.564	270,038 16,940 0.0627 50.661	325,873 20,012 0.0614 47.488	195,796 12,763 0.0652 56.553	350,152 21,454 0.0613 47.152	132,916 9,457 0.0712 70.879	123,388 7,896 0.064 53.691	923,611 38,457 0.0416 0
	1 8,067 852 0.1056 153.654	1 7,657 674 0.088 111.405	1 5,125 720 0.1405 237.406	1 2,382 286 0.1201 188.362	1 5,125 720 0.1405 237.406	1 6,415 858 0.1337 221.221	1 3,769 532 0.1412 239	1 6,750 917 0.1359 226.272	1 2,585 385 0.1489 257.696	1 2,490 330 0.1325 218.294	1 18,137 1,738 0.0958 130.143
	5 9,968 644 0.0646 55.164	3 12,847 777 0.0605 45.256	2 8,569 885 0.1033 148.043	4 3,529 326 0.0924 121.86	2 8,569 885 0.1033 148.043	3 9,744 922 0.0946 127.252	3 6,067 643 0.106 154.537	2 10,187 1,004 0.0986 136.702	5 3,820 420 0.1099 164.058	2 4,037 433 0.1073 157.598	4 25,744 1,595 0.062 48.798
	2 3,326 249 0.0749 79.8	5 3,954 212 0.0536 28.769	5 2,521 231 0.0916 120.066	2 1,004 102 0.1016 143.995	5 2,521 231 0.0916 120.066	5 3,016 276 0.0915 119.782	5 1,860 186 0.1 140.167	3 3,291 310 0.0942 126.229	3 1,141 136 0.1192 186.264	3 1,039 100 0.0962 131.152	3 8,500 541 0.0636 52.858
	11 2,552 133 0.0519 24.677	13 2,004 81 0.0404 -2.926	3 1,290 122 0.0976 134.403	6 629 51 0.0811 94.73	3 1,290 122 0.0976 134.403	4 1,608 151 0.0939 125.53	2 919 103 0.1121 169.175	4 1,669 154 0.0923 121.604	4 656 74 0.113 171.334	4 591 55 0.0931 123.506	10 5,342 252 0.0472 13.295
	3 2,881 206 0.0715 71.727	2 3,242 230 0.0709 70.384	4 2,071 196 0.0946 127.295	3 949 95 0.1001 140.42	4 2,071 196 0.0946 127.295	2 2,614 254 0.0972 133.368	4 1,605 165 0.1028 146.901	5 2,740 239 0.0872 109.489	10 1,036 94 0.0907 117.912	9 958 78 0.0814 95.543	2 7,043 493 0.07 68.114
	6 10,785 649 0.0602 44.523	4 12,768 742 0.0581 39.571	7 8,580 694 0.0809 94.261	7 3,511 283 0.0806 93.584	7 8,580 694 0.0809 94.261	6 9,725 795 0.0817 96.332	6 6,138 550 0.0896 115.204	6 10,670 866 0.0812 94.525	11 3,669 321 0.0875 110.122	5 3,785 339 0.0896 115.104	5 27,331 1,641 0.06 44.2
	10 22,202 1,212 0.0546 31.106	7 23,328 1,200 0.0514 23.543	6 15,593 1,289 0.0827 98.535	5 6,552 533 0.0813 95.374	6 15,593 1,289 0.0827 98.535	7 17,652 1,376 0.078 87.214	8 10,797 915 0.0847 103.532	7 19,050 1,522 0.0799 91.882	9 6,639 604 0.091 118.498	7 6,435 552 0.0858 106.018	6 52,978 2,786 0.0526 26.299
	22 26,685 1,160 0.0435 4.401	10 35,405 1,530 0.0432 3.786	8 19,832 1,572 0.0793 90.371	9 7,844 552 0.0704 69.011	8 19,832 1,572 0.0793 90.371	8 21,743 1,641 0.0755 81.26	7 13,721 1,167 0.0851 104.267	8 22,168 1,743 0.0786 88.836	8 8,249 788 0.0955 129.424	8 8,327 689 0.0827 98.721	18 74,559 3,167 0.0425 2.014
	4 7,745 518 0.0669 60.628	26 7,202 185 0.0257 -38.308	13 4,898 322 0.0657 57.889	15 2,308 148 0.0641 54.007	13 4,898 322 0.0657 57.889	11 6,051 423 0.0699 67.891	19 3,652 235 0.0643 54.544	9 6,436 506 0.0786 88.82	2 2,562 308 0.1202 188.726	12 2,369 169 0.0716 72.057	7 17,235 834 0.0484 16.217
	7 7,699 460 0.0597 43.495	29 7,201 169 0.0235 -43.635	11 4,809 340 0.0707 69.8	10 2,269 158 0.0696 67.239	11 4,809 340 0.0707 69.8	9 6,004 433 0.0721 73.205	14 3,544 243 0.0686 64.674	10 6,247 475 0.076 82.615	6 2,482 257 0.1035 148.682	11 2,329 167 0.0717 72.211	12 17,169 783 0.0456 9.529
	12 7,688 393 0.0511 22.77	9 7,229 336 0.0465 11.528	9 4,785 363 0.0759 82.196	17 2,280 139 0.061 46.418	9 4,785 363 0.0759 82.196	12 6,037 403 0.0668 60.324	13 3,501 245 0.07 68.069	11 6,319 430 0.068 63.431	15 2,397 182 0.0759 82.355	15 2,312 152 0.0657 57.895	8 17,275 833 0.0482 15.809
	8 2,138 126 0.0589 41.539	14 2,479 95 0.0383 -7.963	10 1,546 117 0.0757 81.757	8 733 56 0.0764 83.484	10 1,546 117 0.0757 81.757	10 1,941 137 0.0706 69.515	9 1,206 91 0.0755 81.221	12 2,026 137 0.0676 62.403	13 723 58 0.0802 92.665	6 762 66 0.0866 108.019	11 5,347 252 0.0471 13.189
	24 28,932 1,160 0.0401	31 11,695 262 0.0224	15 11,432 736 0.0644	24 6,301 348 0.0552	15 11,432 736 0.0644	19 14,567 877 0.0602	20 8,011 502 0.0627	13 16,878 1,085 0.0643	12 7,066 577 0.0817	24 5,444 293 0.0538	25 45,041 1,623 0.036

Examples

- **ACQUISITION:** A “Star Trek” package was #3 with 18-20 demo, #2 with 21-24 demo, but #9 overall. We can acquire younger audiences with targeted content like this.



- **ENGAGEMENT:** “Kobe’s astonishing shot” was #25 with women, but #5 with men. We can better engage men (or sports fans) by showing more like this, women by showing less.



- **REACH:** A package about a hair-pulling soccer player was just plain interesting to everyone (#1-3). We can maintain reach by programming content for the mass audience.

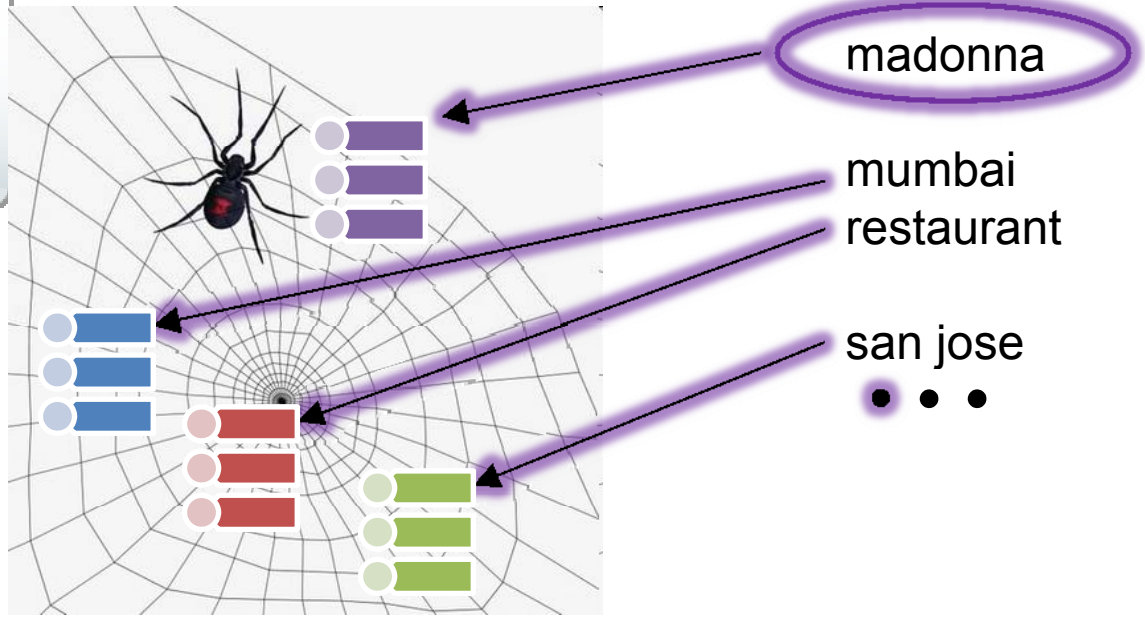
NEXT-GEN SEARCH


berts - Yahoo! Search Results - Mozilla Firefox
View History Bookmarks Yahoo! Tools Help
http://search.yahoo.com/search?p=julia+roberts&ei=UTF-8&fr=moz35
julia roberts - Yahoo! Search Res...
YAHOO!
julia roberts
Search
julia roberts twins
julia roberts movies
lyle julia roberts
julia roberts babies
julia roberts henry daniel moder
Explore related concepts:
actor
julia roberts photos
Julia Roberts News
Search Pad
SearchScan - On
40,500,000 results for
julia roberts
Related People
Scarlett Johansson
Emma Robe
Hila Swank
Lindsay Lohan
Tom Hanks
Halle Berry
Julia Roberts - Image Result
News & Photos
Twitter
Hindus concerned about Julia Roberts' "Eat, Pray, Love" - New Kerala - 6 hours ago
Trailer For 'Eat Pray Love' Starring Julia Roberts - KPBS San Diego - Mar 19 03:18pm
Link Party: Julia Roberts' New Movie Will Teach You How to Live - E! Online - Mar 18 05:48pm
more Julia Roberts news...
Julia Roberts - Wikipedia
Early life | Career | Influence | Personal life
Julia Fiona Roberts is an American actress. She is known for starring in the romantic comedy Pretty Woman opposite Richard Gere, which grossed $463 million worldwide. After receiving...
en.wikipedia.org/wiki/Julia_Roberts - 122k - Cached
Concept-centric Information aggregation

Web of Concepts

rich, aggregated data

concept

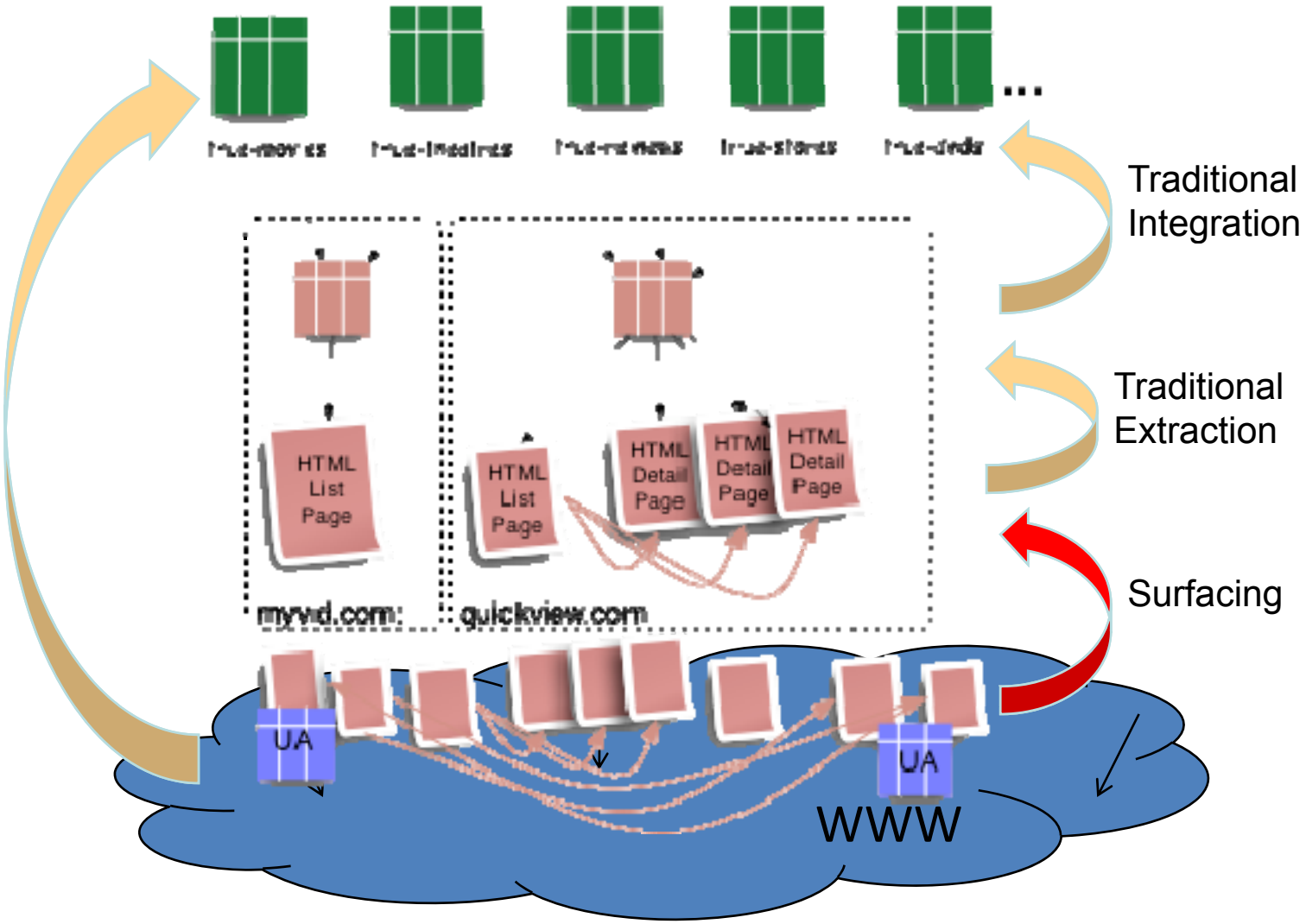


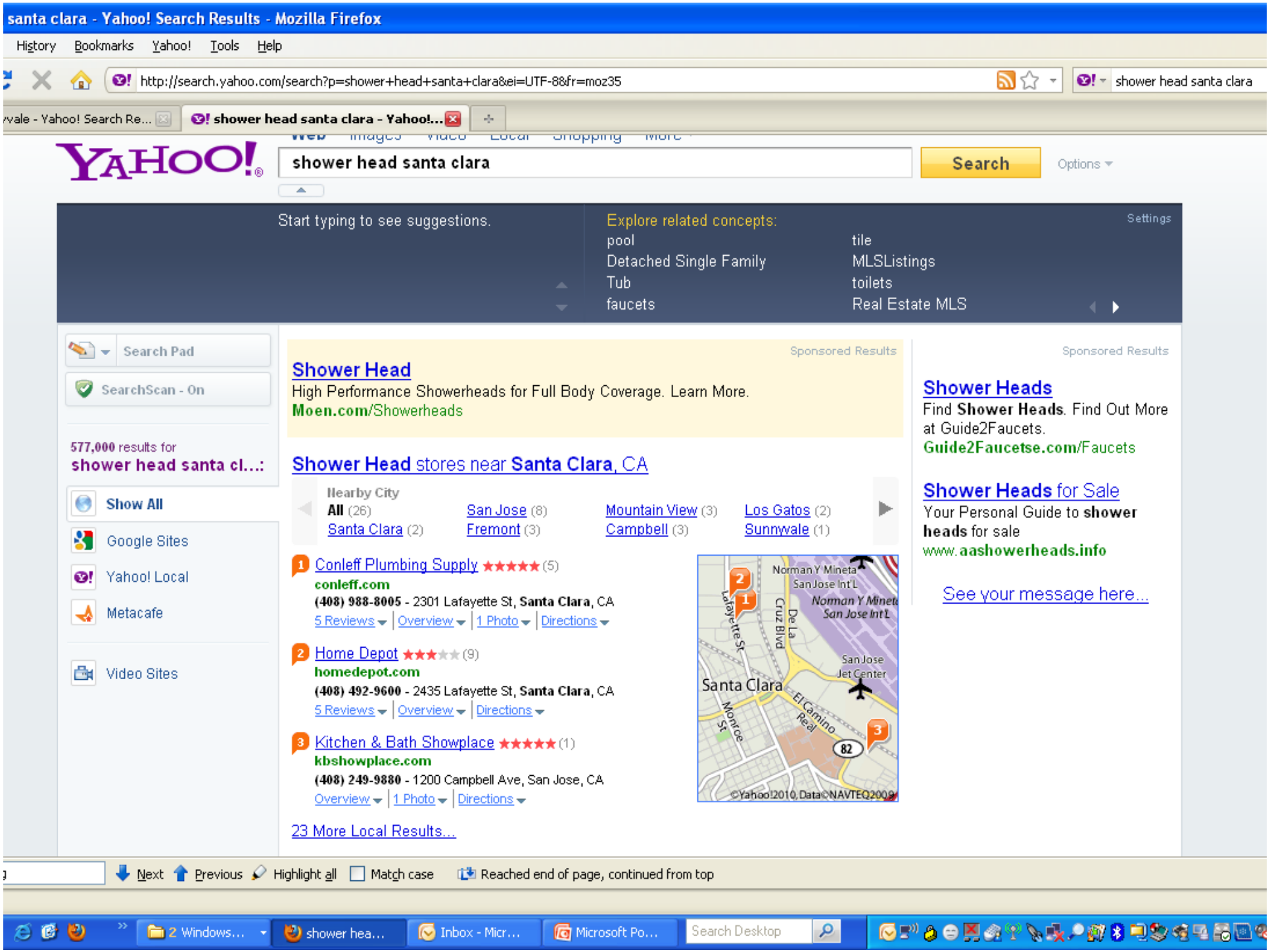
Aggregated KB **INDEX** **SERP**

The “index” is keyed by concept instance, and organizes all relevant information (data describing the concept instance and its relationship to other instances), wherever it is drawn from, in semantically meaningful ways

Web IE: Surfacing, Extraction, Integration

End-to-End




santa clara - Yahoo! Search Results - Mozilla Firefox
History Bookmarks Yahoo! Tools Help
http://search.yahoo.com/search?p=shower+head+santa+clara&ei=UTF-8&fr=moz35
shower head santa clara
shower head santa clara - Yahoo!...
YAHOO!
shower head santa clara
Search
Options
Start typing to see suggestions.
Explore related concepts:
pool
Detached Single Family
Tub
faucets
tile
MLSListings
toilets
Real Estate MLS
Settings
Search Pad
SearchScan - On
577,000 results for
shower head santa cl...:
Show All
Google Sites
Yahoo! Local
Metacafe
Video Sites
Sponsored Results
Shower Head
High Performance Showerheads for Full Body Coverage. Learn More.
Moen.com/Showerheads
Shower Head stores near Santa Clara, CA
Nearby City
All (26)
Santa Clara (2)
San Jose (8)
Fremont (3)
Mountain View (3)
Campbell (3)
Los Gatos (2)
Sunnyvale (1)
1 Conleff Plumbing Supply ★★★★★ (5)
conleff.com
(408) 988-8005 - 2301 Lafayette St, Santa Clara, CA
5 Reviews | Overview | 1 Photo | Directions
2 Home Depot ★★★☆☆ (9)
homedepot.com
(408) 492-9600 - 2435 Lafayette St, Santa Clara, CA
5 Reviews | Overview | Directions
3 Kitchen & Bath Showplace ★★★★★ (1)
kbshowplace.com
(408) 249-9880 - 1200 Campbell Ave, San Jose, CA
Overview | 1 Photo | Directions
23 More Local Results...
Norman Y Mineta San Jose Int'l
Lafayette St
De La Cruz Blvd
San Jose Jet Center
Santa Clara
El Camino Real
Monroe St
82
©Yahoo!2010, Data©NAVTEQ2009
Sponsored Results
Shower Heads
Find Shower Heads. Find Out More at Guide2Faucets.
Guide2Faucetse.com/Faucets
Shower Heads for Sale
Your Personal Guide to shower heads for sale
www.aashowerheads.info
See your message here...
Next Previous Highlight all Match case Reached end of page, continued from top
2 Windows... shower hea... Inbox - Micr... Microsoft Po... Search Desktop

ry Bookmarks Yahoo! Tools Help

http://search.yahoo.com/search;_ylt=A0oGklHEA_RLAyYBgxJXNyoA?p=eggplant+parmigiana+baltimore&fr2=sb-top&fr=moz35&sao=1

hummus

na baltimore - Yah...

YAHOO!

Web Images Video Local Shopping News More

eggplant parmigiana baltimore

Search

Options

Start typing to see suggestions.

Explore related concepts:

Pizza	mushrooms
veal	tomato sauce
Chicken	Little Italy
Italian Restaurant	Crab Cake

Search Pad

SearchScan - On

30,300 results for
eggplant parmigiana ...:

Show All

Los Angeles Times

Local Business Sites

Eggplant Parmigiana Restaurants near Baltimore, MD

Neighborhood

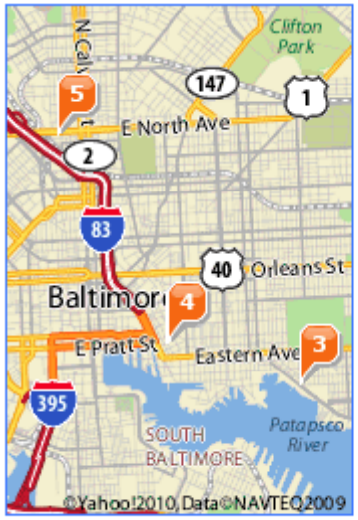
All (36)	Canton (2)	Charles North (1)	Downtown (1)
Abell (1)	Central Bal... (4)	Chinquapin ... (1)	Federal Hill (1)

1 Ciao Bella - Baltimore ★★★★★ (11)
local.yahoo.com
(410) 685-7733 - 236 S High St, Baltimore, MD
Menu: eggplant parmigiana
4 Reviews | Overview | 2 Photos | Directions

2 Amicci's ★★★★★ (20)
amiccis.com
(410) 528-1096 - 231 S High St, Baltimore, MD
Menu: eggplant parmigiana
14 Reviews | Overview | 23 Photos | Directions

3 Pasticcio ★★★★☆ (8)
local.yahoo.com
(410) 522-7700 - 2400 Boston St, Baltimore, MD
Menu: eggplant parmigiana
5 Reviews | Overview | 3 Photos | Directions

4 Caesar's Den ★★★★☆ (7)
caesarsden.com
(410) 547-0820 - 223 S High St, Baltimore, MD
Menu: eggplant parmigiana
4 Reviews | Overview | 11 Photos | Directions





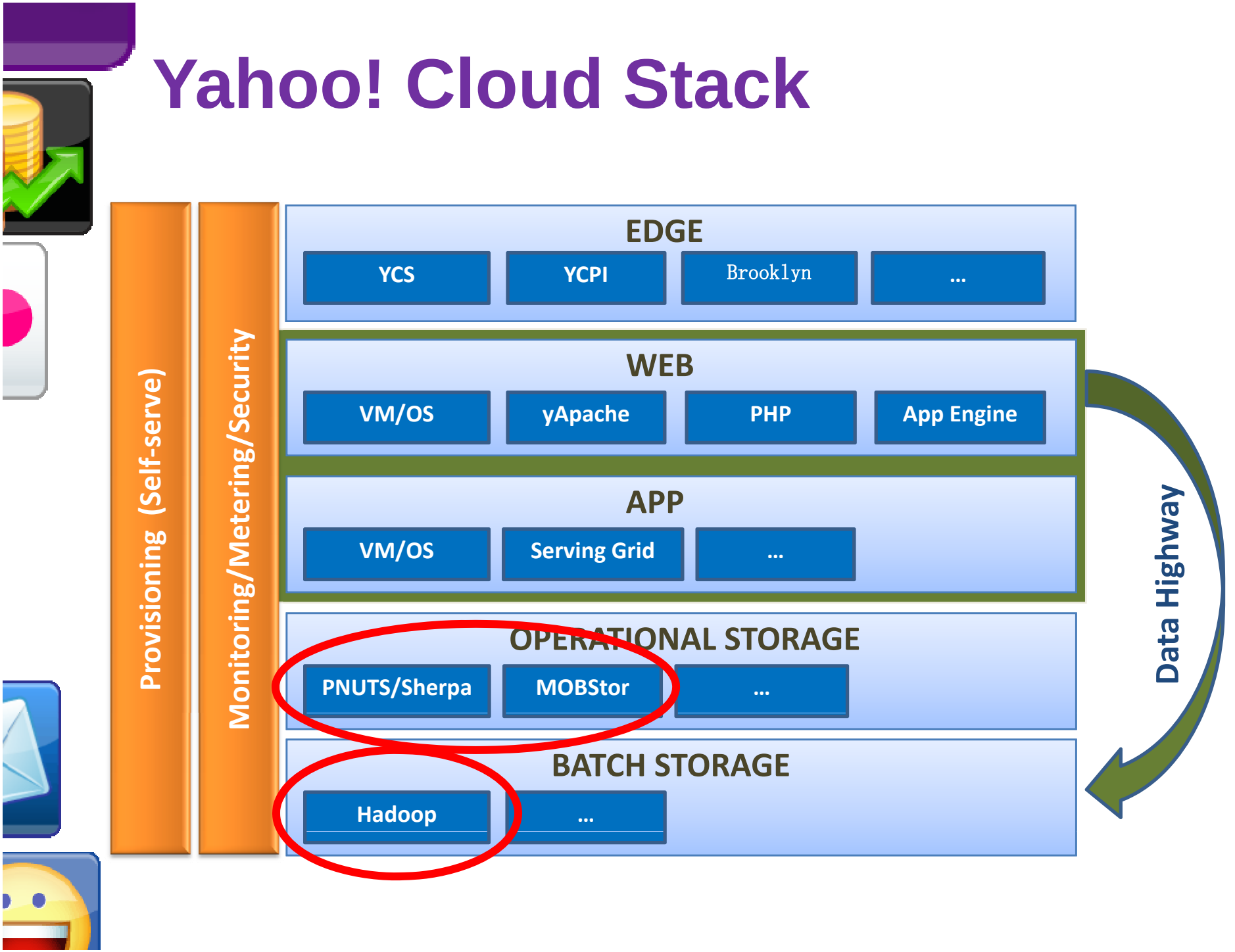


DATA MANAGEMENT IN THE CLOUD

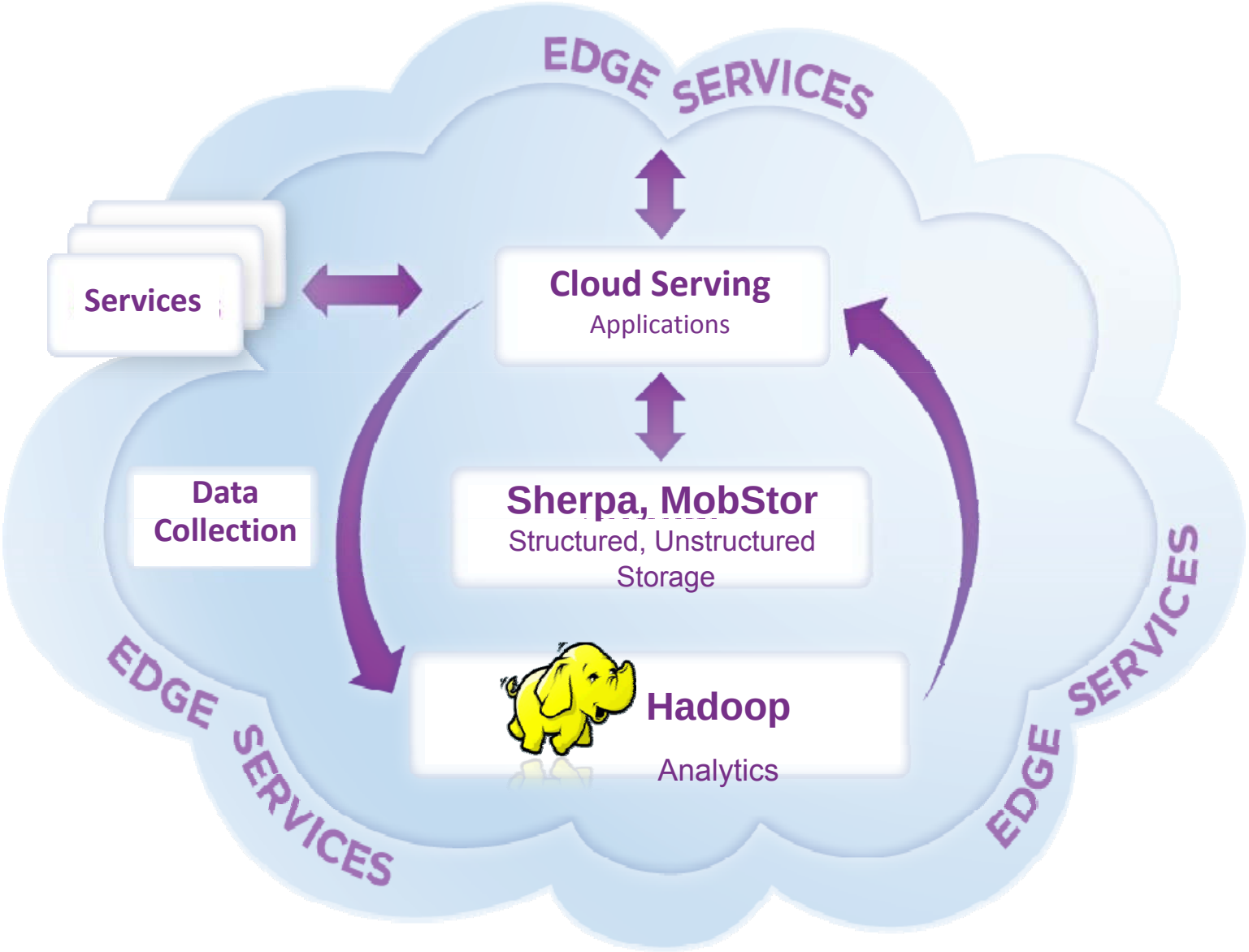
Requirements for Cloud Services

- **Multitenant.** A cloud service must support multiple, organizationally distant customers.
- **Elasticity.** Tenants should be able to negotiate and receive resources/QoS *on-demand* up to a large scale.
- **Resource Sharing.** Ideally, spare cloud resources should be transparently applied when a tenant's negotiated QoS is insufficient, e.g., due to spikes.
- **Horizontal scaling.** The cloud provider should be able to add cloud capacity in increments without affecting tenants of the service.
- **Metering.** A cloud service must support accounting that reasonably ascribes operational and capital expenditures to each of the tenants of the service.
- **Security.** A cloud service should be secure in that tenants are not made vulnerable because of loopholes in the cloud.
- **Availability.** A cloud service should be highly available.
- **Operability.** A cloud service should be easy to operate, with few operators. Operating costs should scale linearly or better with the capacity of the service.

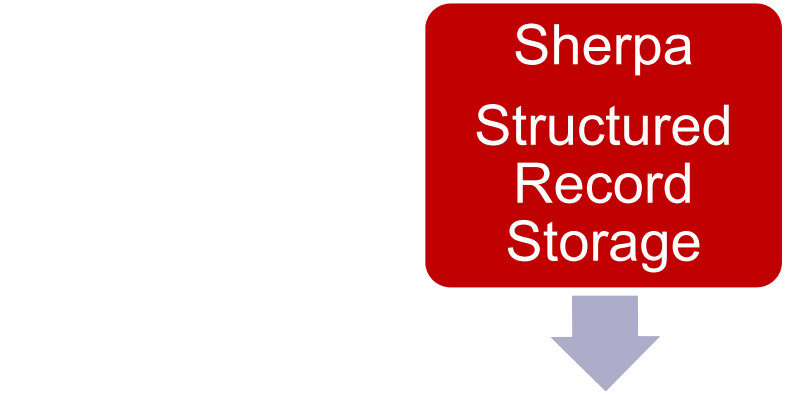
Yahoo! Cloud Stack



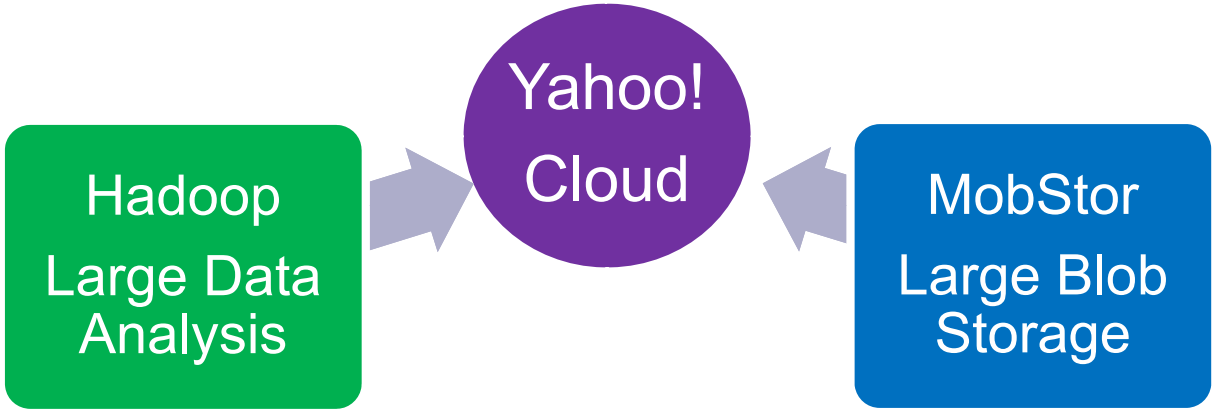
A Data-Centric View



Cloud Data Management



- CRUD
- Point lookups and short scans
- Index organized table and random I/Os

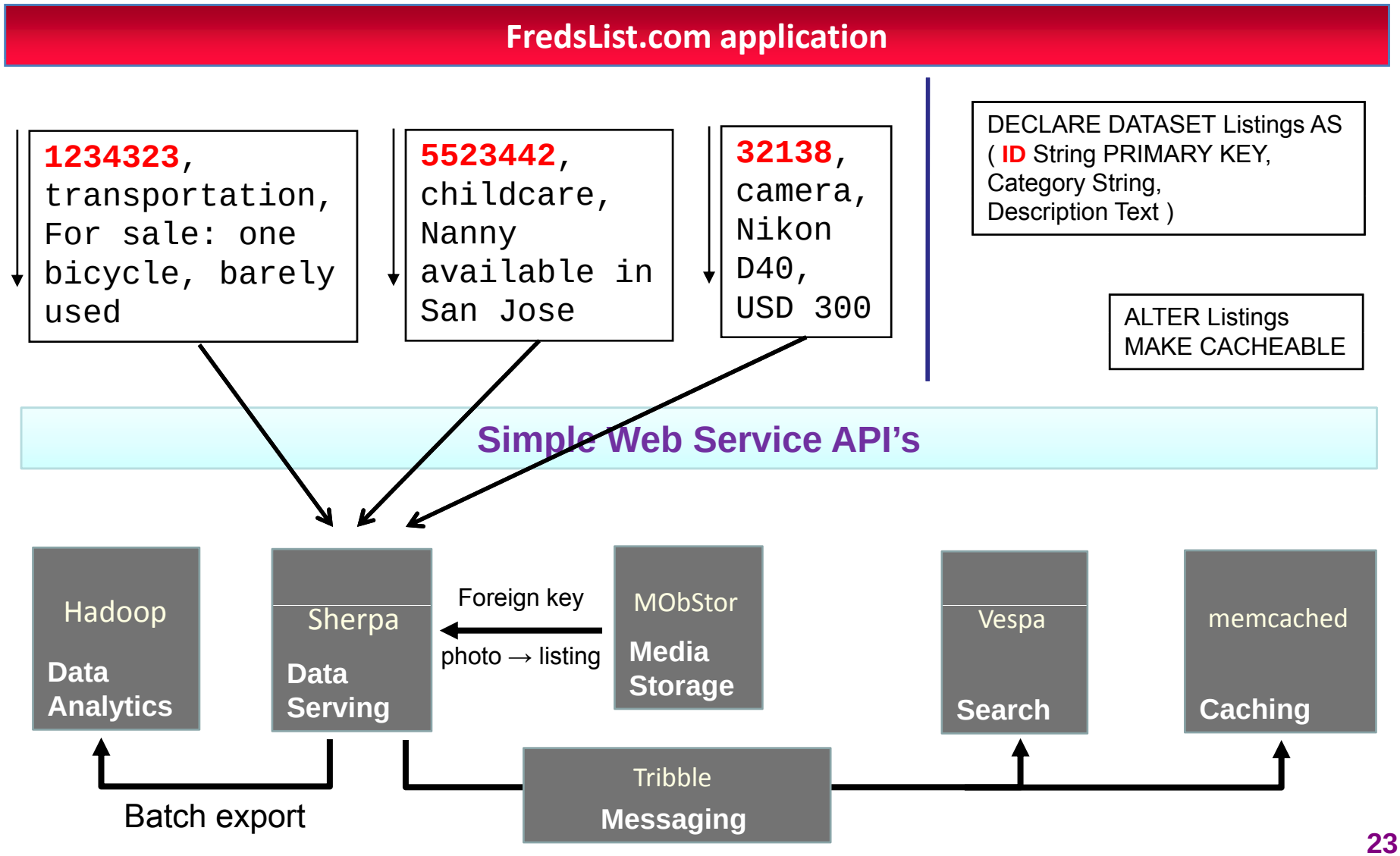


- Scan oriented workloads
- Focus on Sequential disk I/O

- Object retrieval and streaming
- Scalable file storage

22

The Yahoo! Data Cloud





Hadoop Core (Core, Pig, Oozie, Hive, Howl)	Ad BT and Inventory prediction, Content Agility, UDA, COKE, Mail Spam, Search, APG, Labs, Insights, Analytics	1+ million jobs per month 3.7 PB processed daily 90B events and 120 TB daily 70+ PB of Data

Map-Reduce and more …

HADOOP: SCALABLE ANALYTICS


One Slide Hadoop Primer

Good for analyzing (scanning) huge files
Not great for serving (reading or writing individual objects)

Industry Challenge: Massive Data Sets Being Accumulated

"It's now the industrial revolution of data" – Joe Hellerstein, UC Berkeley

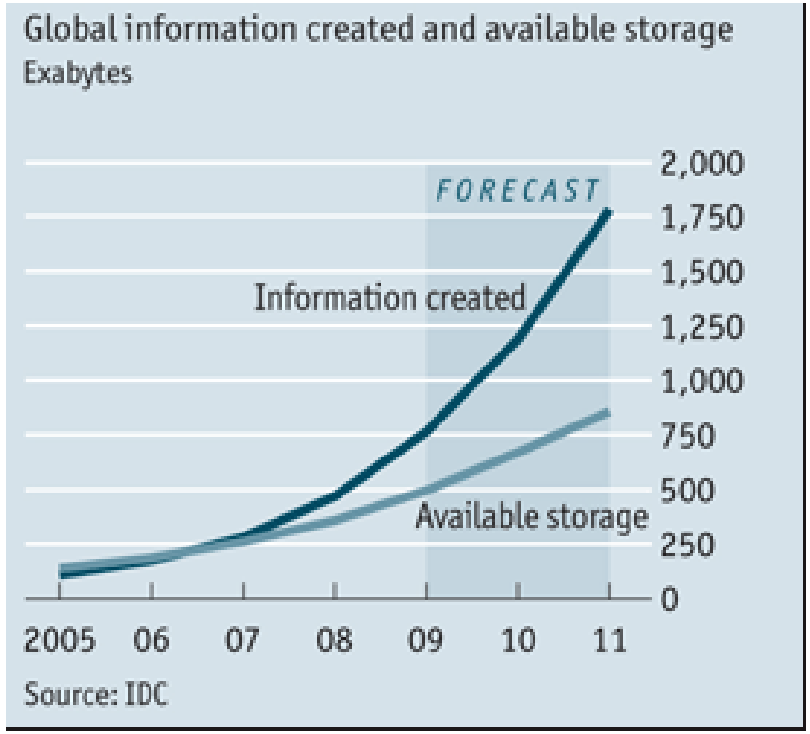
1. Massive datasets of highly-leveragable data amassed by relatively few players

"By 2020, the Digital Universe will be 44 times as large as it was in 2009" [IDC]

2. Deriving insights from the newly available mountains of data still challenging

3. Datasets become competitive differentiators

Search query history, Advertising/click-through data, Surfing histories, Social graphs, News, Finance, Sports and other data feeds



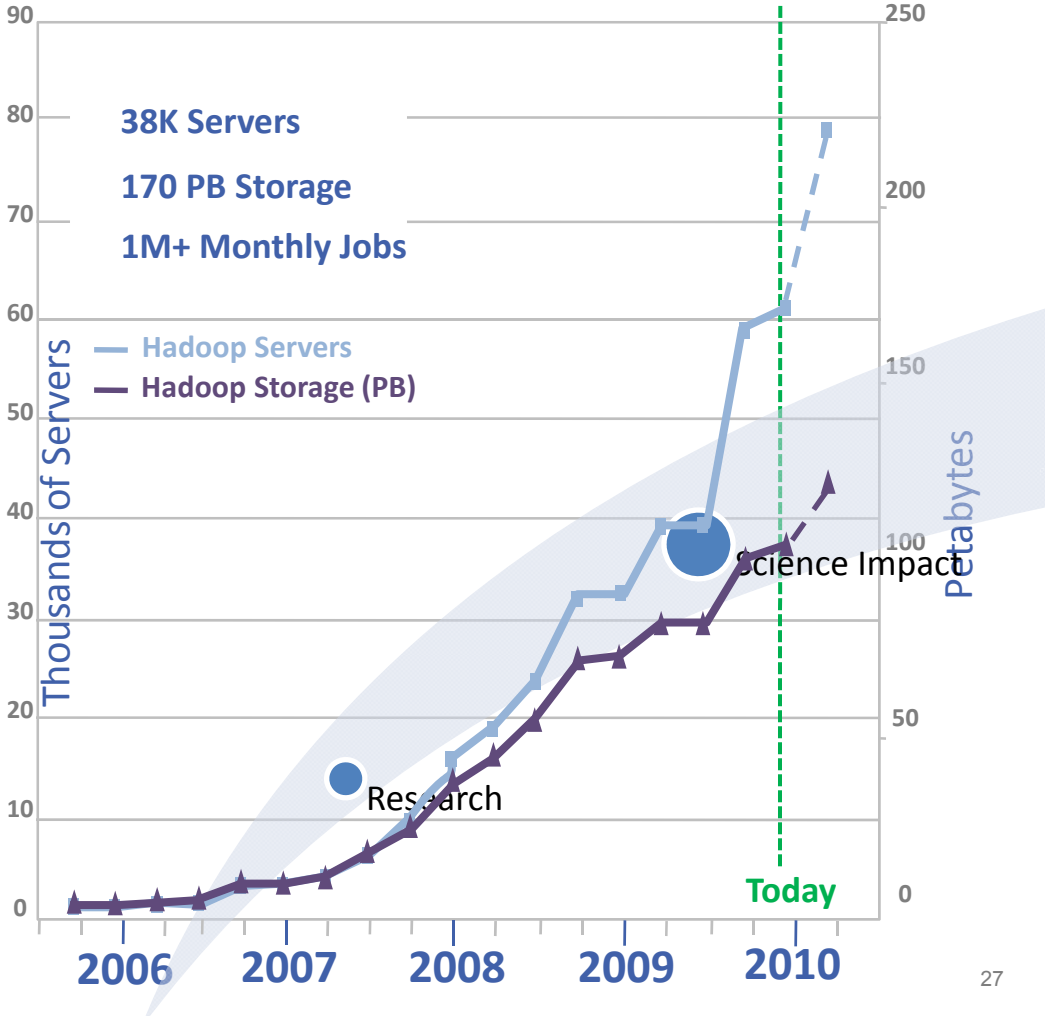
Yahoo!'s technology solution: Hadoop

Hadoop: Stability at Scale

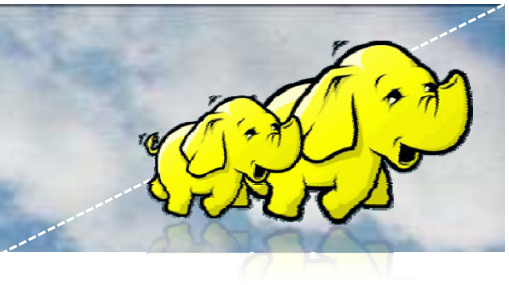
Hadoop powers the Yahoo! Network: must be rock-solid

We fix bugs before you see them
- We run very large clusters
- We have a large QA effort
- We run a huge variety of workloads

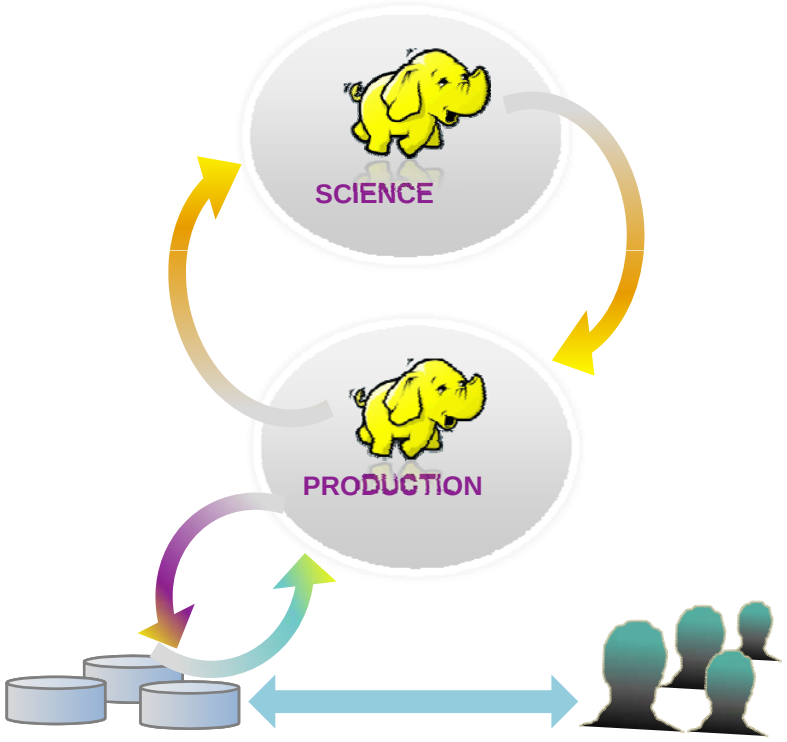
The Yahoo! Distribution of Hadoop
- We contribute our work to Apache
- We share the exact code we run
- We don't sell software or service



CASE STUDY
YAHOO! MAIL



Enabling quick response in the spam arms race



- **450M mail boxes**
- **5B+ deliveries/day**

- **Antispam models retrained every few hours on Hadoop**

> **“40% less spam than Hotmail and 55% less spam than Gmail”**

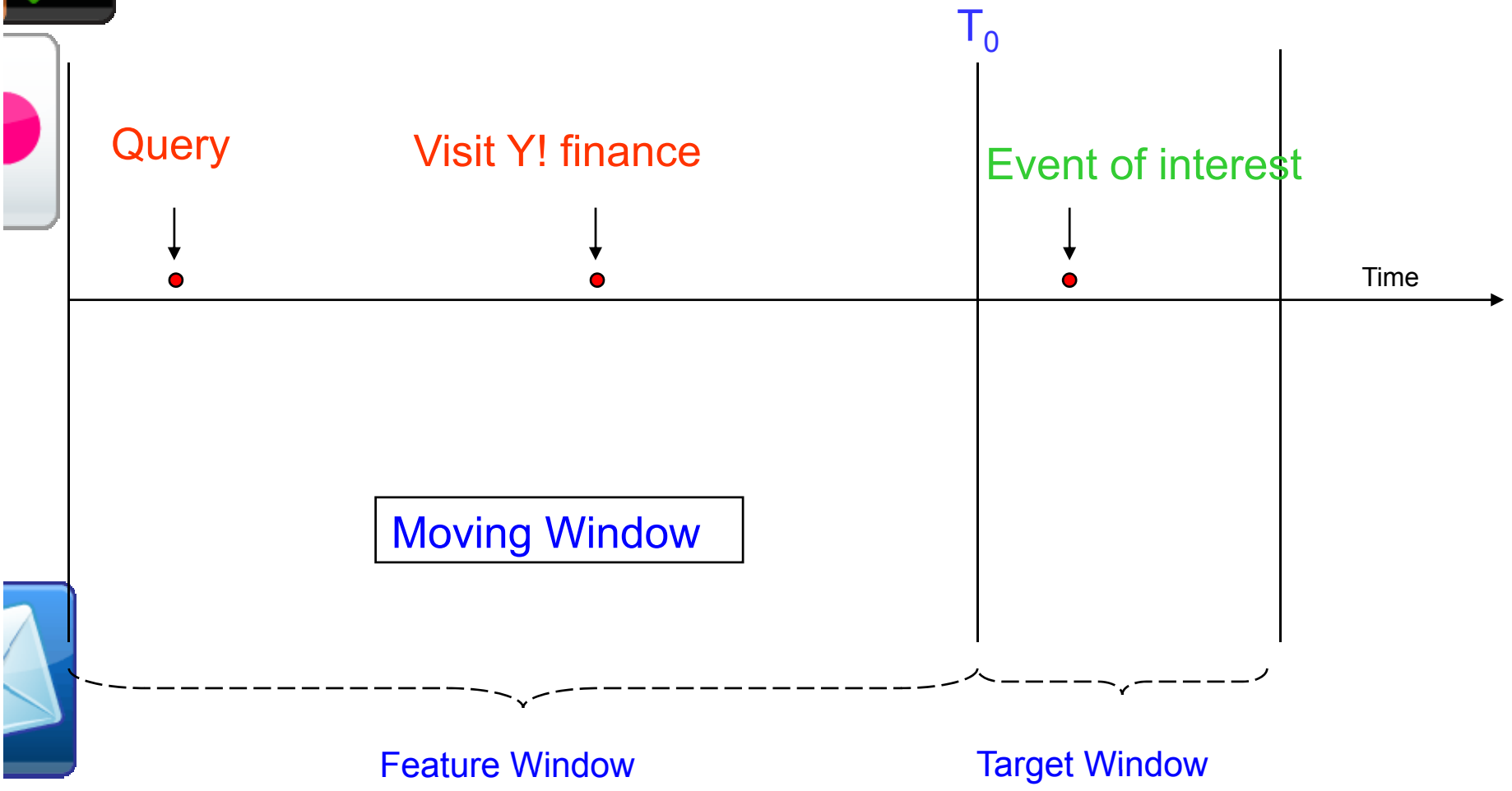
Example: User Activity Modeling

Large dimensionality vector describing possible user activities
But a typical user has a sparse activity vector

Attribute	Possible Values	Typical values per user
Pages	~ MM	10 – 100
Queries	~ 100s of MM	Few
Ads	~ 100s of thousands	10s

Hadoop pipeline to model user interests from activities

Feature and Target Windows



User Modeling Pipeline

Component	Data Processed	Time
Data Acquisition	~ 1 Tb per time period	2 – 3 hours
Feature and Target Generation	~ 1 Tb * Size of feature window	4 - 6 hours
Model Training	~ 50 - 100 Gb	1 – 2 hours for 100's of models
Scoring	~ 500 Gb	1 hour

A Growing User Base

Year: 2007

Year: 2008

Year: 2009 - 2010

Y!OS, COKE, LocDrop, Video, Media
Search history, Answers, Messenger,
BOSS, Image Search, Blog Search

15K requests per second
Over 1.5B records; 10sTB of data

ACID or BASE? Litmus tests are colorful, but the picture is cloudy

PNUTS:
SCALABLE DATA SERVING

Requirements for Cloud Services

- **Multitenant.** A cloud service must support multiple, organizationally distant customers.
- **Elasticity.** Tenants should be able to negotiate and receive resources/QoS *on-demand* up to a large scale.
- **Resource Sharing.** Ideally, spare cloud resources should be transparently applied when a tenant's negotiated QoS is insufficient, e.g., due to spikes.
- **Horizontal scaling.** The cloud provider should be able to add cloud capacity in increments without affecting tenants of the service.
- **Metering.** A cloud service must support accounting that reasonably ascribes operational and capital expenditures to each of the tenants of the service.
- **Security.** A cloud service should be secure in that tenants are not made vulnerable because of loopholes in the cloud.
- **Availability.** A cloud service should be highly available.
- **Operability.** A cloud service should be easy to operate, with few operators. Operating costs should scale linearly or better with the capacity of the service.

The World Has Changed

Web serving applications need:

- Scalability!
  - Elastic on demand, commodity boxes
- Flexible schemas
- Geographic distribution/replication
- High availability
- Low latency

Web serving applications willing to do without:

- Complex queries
- ACID transactions
  - But still benefit from support for data consistency



Typical Y! Applications

User logins and profiles

- Including changes that must not be lost!
  - But single-record “transactions” suffice

Events

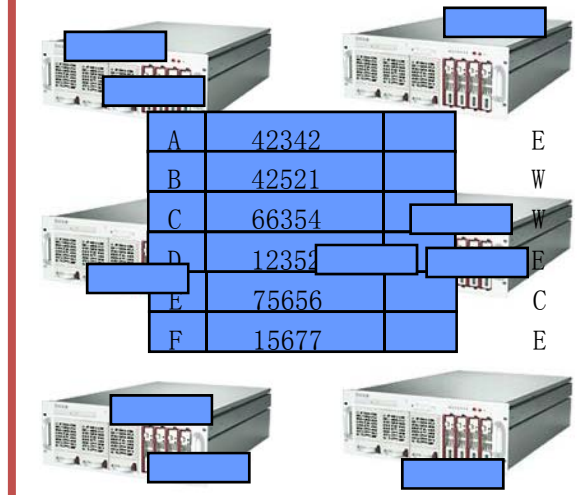
- Alerts (e.g., news, price drops)
- Social network activity (e.g., user goes offline)
- Ad clicks, article clicks

Application-specific data

- Postings in message board
- Uploaded photos, tags
- Shopping carts


640M+ unique users, 11B pages/month
Hundreds of petabytes of storage
Hundreds of billions of objects
Hundred of thousands of requests/sec
Global, rapidly evolving workloads

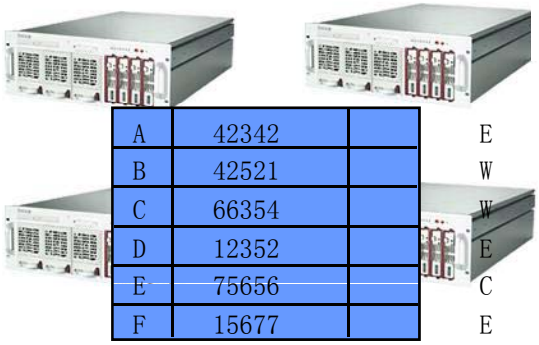
What is PNUTS/Sherpa?



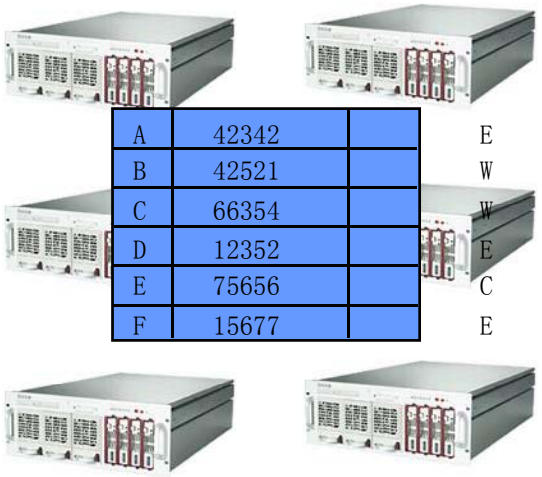
Parallel database

```
CREATE TABLE Parts (
    ID VARCHAR,
    StockNumber INT,
    Status VARCHAR
    ...
)
```

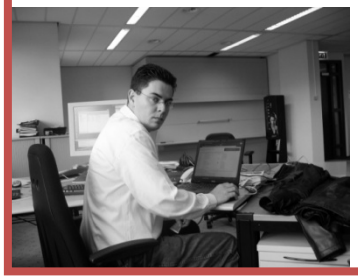
Structured, flexible schema



Hosted, managed infrastructure



Geographic replication



PNUTS: Key Components

- Caches the maps from the TC
- Routes client requests to correct SU

VIP

Routers

- Maintains map from database.table.key-to-tablet-to-SU
- Provides load balancing

Tablet Controller

Table: FOO

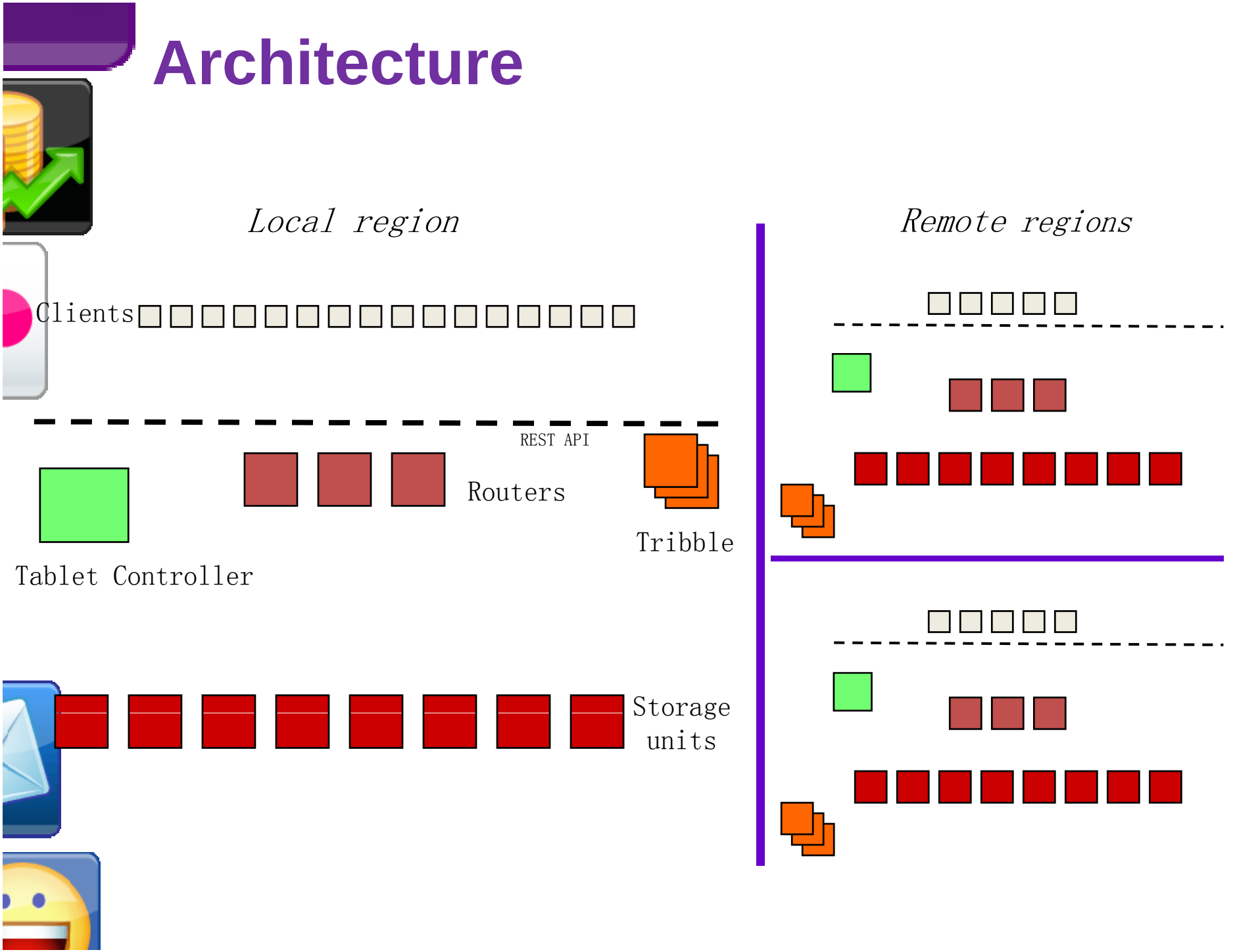
	Key	JSON	
1			*Tablet 1*
3			*Tablet 2*
5			*Tablet 3*
2			*Tablet 4*
9			*Tablet 5*
n			*Tablet M*

1 2 n

Storage Units

- Stores records in **tablets**
- Services get/set/delete requests

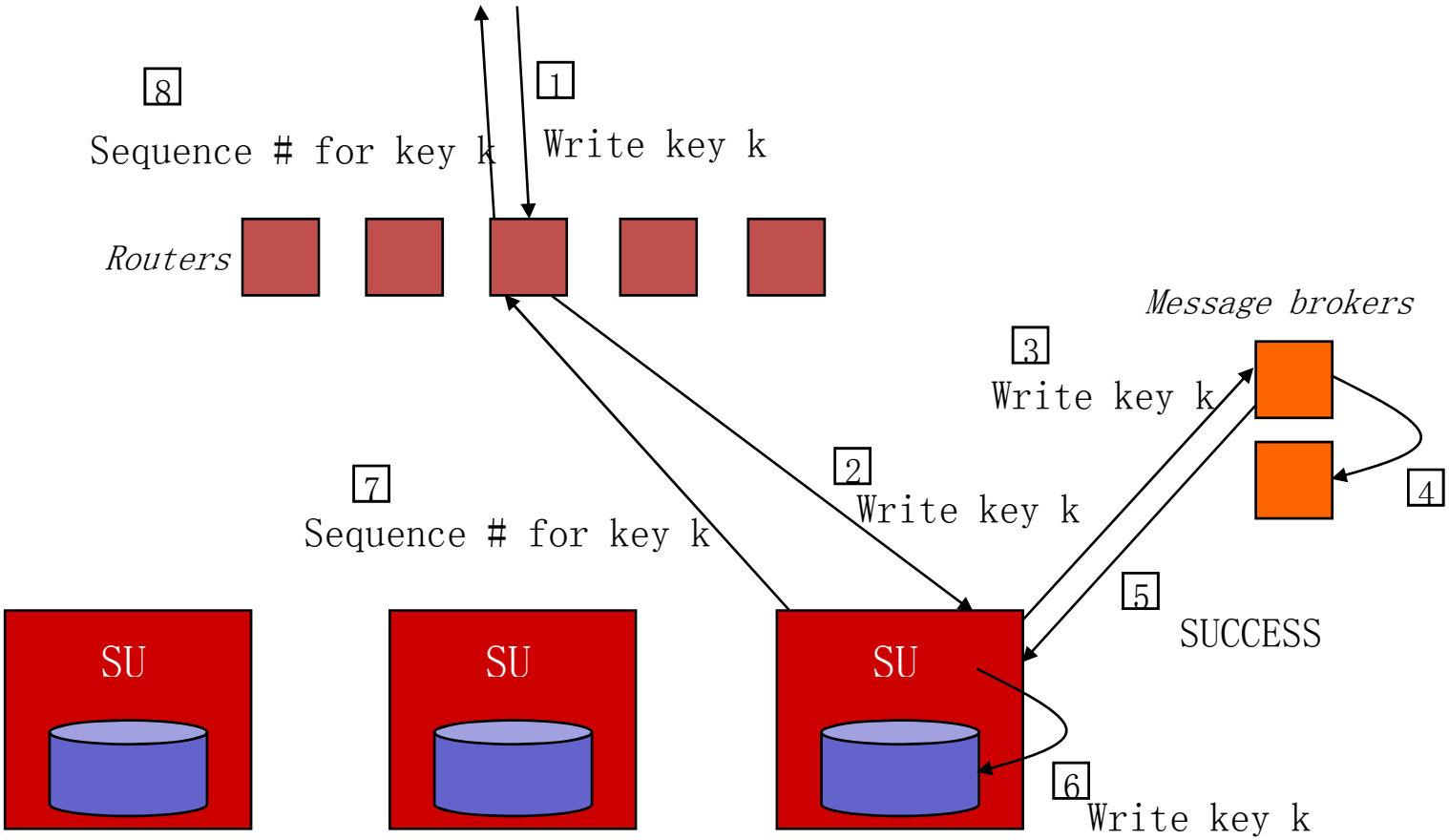
Architecture



Flexible Schema

Posted date	*Listing id*	*Item*	*Price*	*Color*	*Condition*
6/1/07	424252	Couch	$570		Good
6/1/07	763245	Bike	$86		
6/3/07	211242	Car	$1123	Red	Fair
6/5/07	421133	Lamp	$15		

Updates


8
Sequence # for key k
1
Write key k
Routers
Message brokers
3
Write key k
2
Write key k
4
7
Sequence # for key k
5
SUCCESS
SU
SU
SU
6
Write key k

Tablets—Ordered Table





A

H

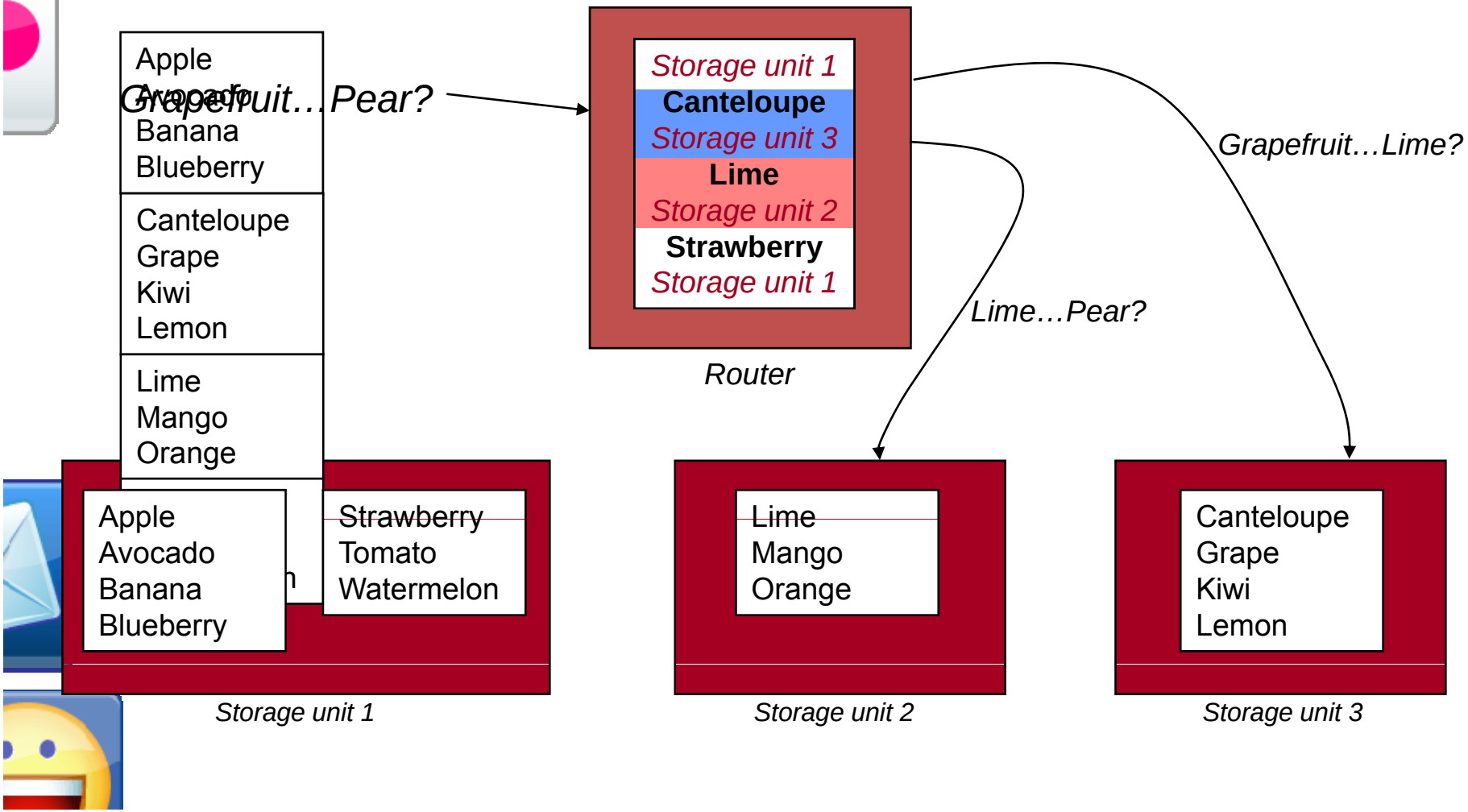
Q

Z

Name	*Description*	*Price*
Apple	Apple is wisdom	$1
Avocado	But at what price?	$3
Banana	The perfect fruit	$2
Grape	Grapes are good to eat	$12
Kiwi	New Zealand	$8
Lemon	How much did you pay for this lemon?	$1
Lime	Limes are green	$9
Orange	Arrgh! Don’t get scurvy!	$2
Strawberry	Strawberry shortcake	$900
Tomato	Is this a vegetable?	$14

Range Queries in YDOT

Clustered, ordered retrieval of records

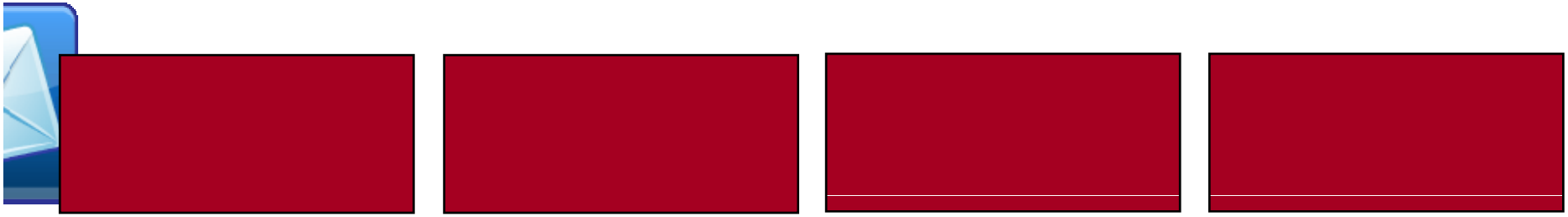


ELASTICITY, OPERABILITY, HORIZONTAL SCALING

Distribution

Data shuffling for load balancing

6/1/07			570
6/1/07	256623	Car	$1123
6/2/07	636353	Bike	$86
6/5/07	662113	Chair	$10
6/7/07	121113	Lamp	$19
6/9/07	887734	Bike	$56
6/11/07	252111	Scooter	$18
6/11/07	116458	Hammer	$8000

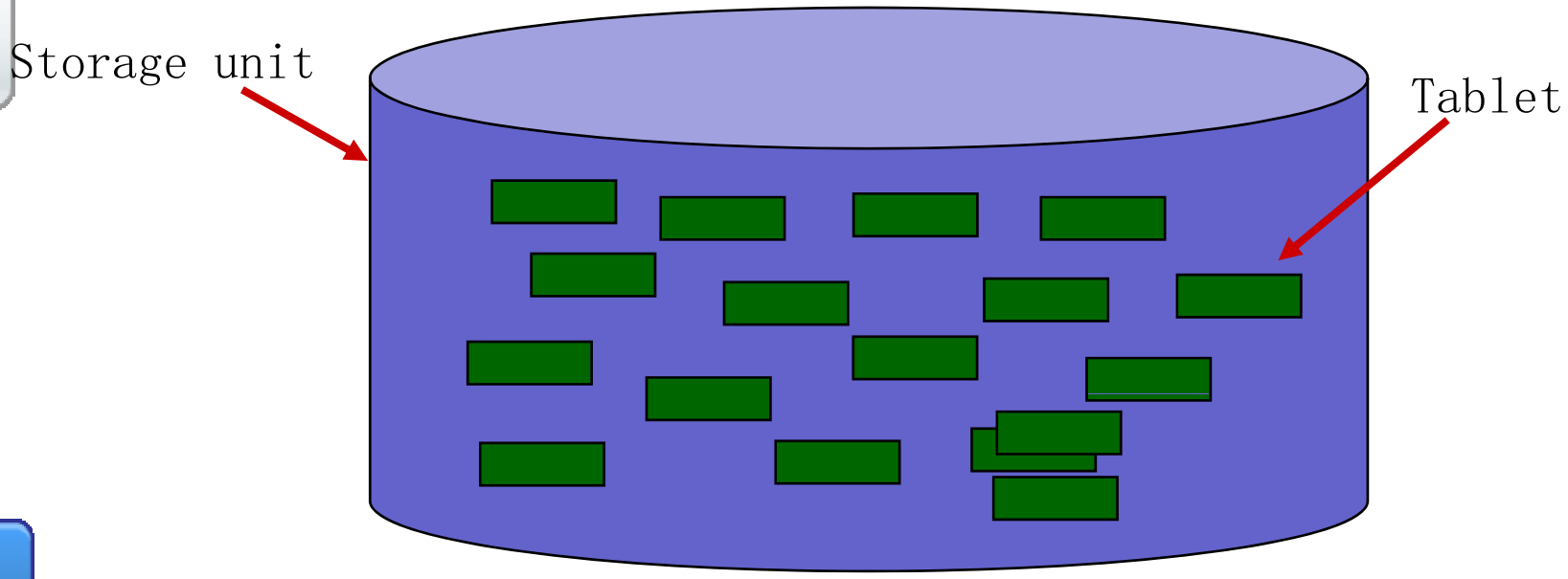


Server 1 *Server 2* *Server 3* *Server 4*

Tablet Splitting and Balancing

Each storage unit has many tablets (horizontal partitions of the table)

Storage unit may become a hotspot



Overfull tablets split

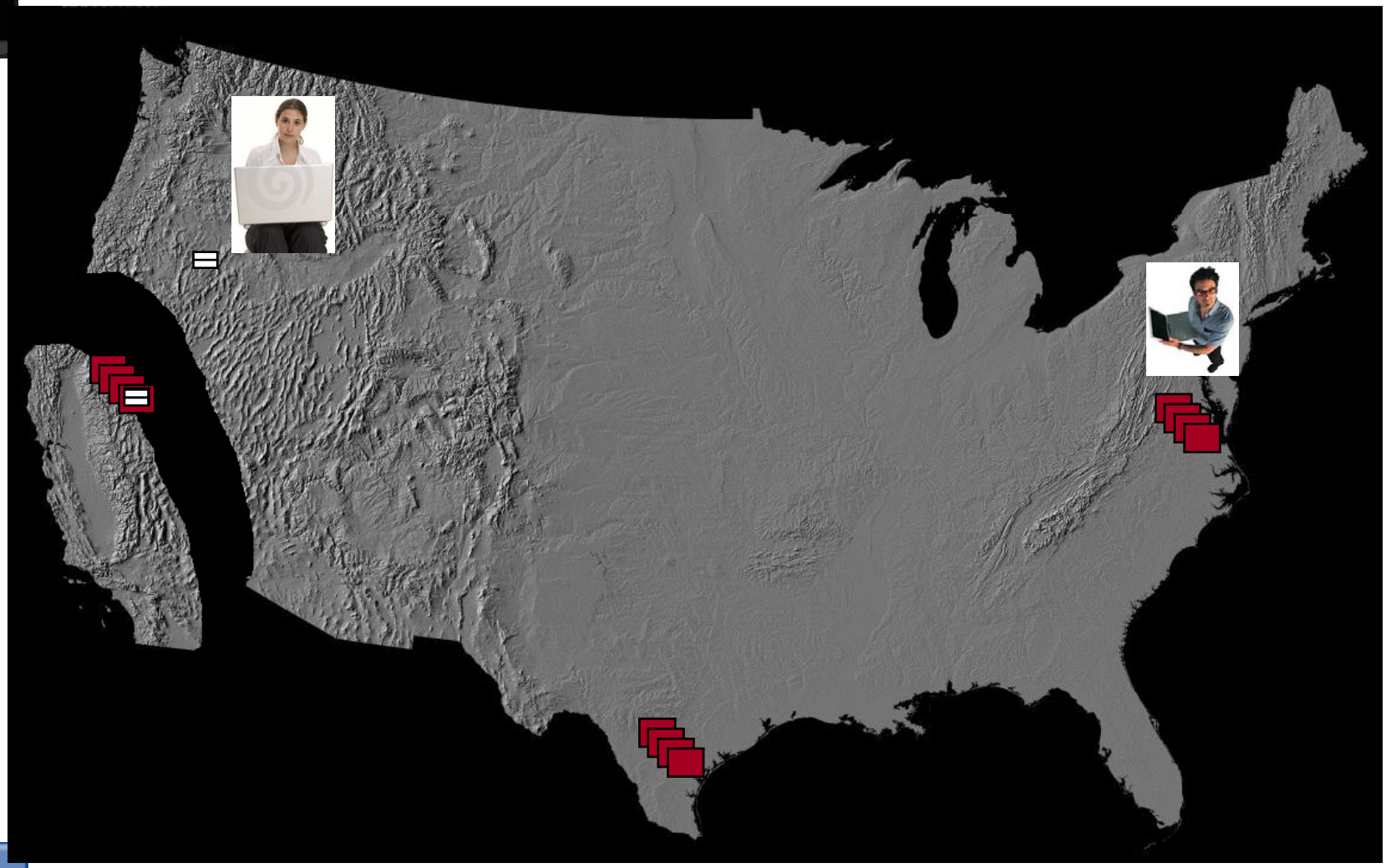
Tablets may grow over time

Shed load by moving tablets to other servers



ASYNCHRONOUS REPLICATION AND CONSISTENCY


Asynchronous Replication



Consistency: Social Alice

West

User	Status
Alice	Busy

Network disruption:
Alice redirected to East

User	Status
Alice	Free

User	Status
Alice	Busy

busy

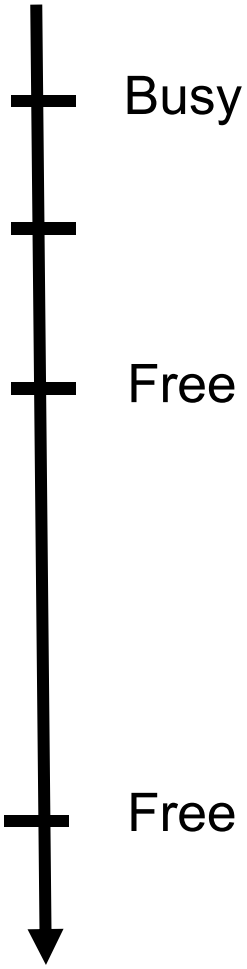
free

User	Status
Alice	???

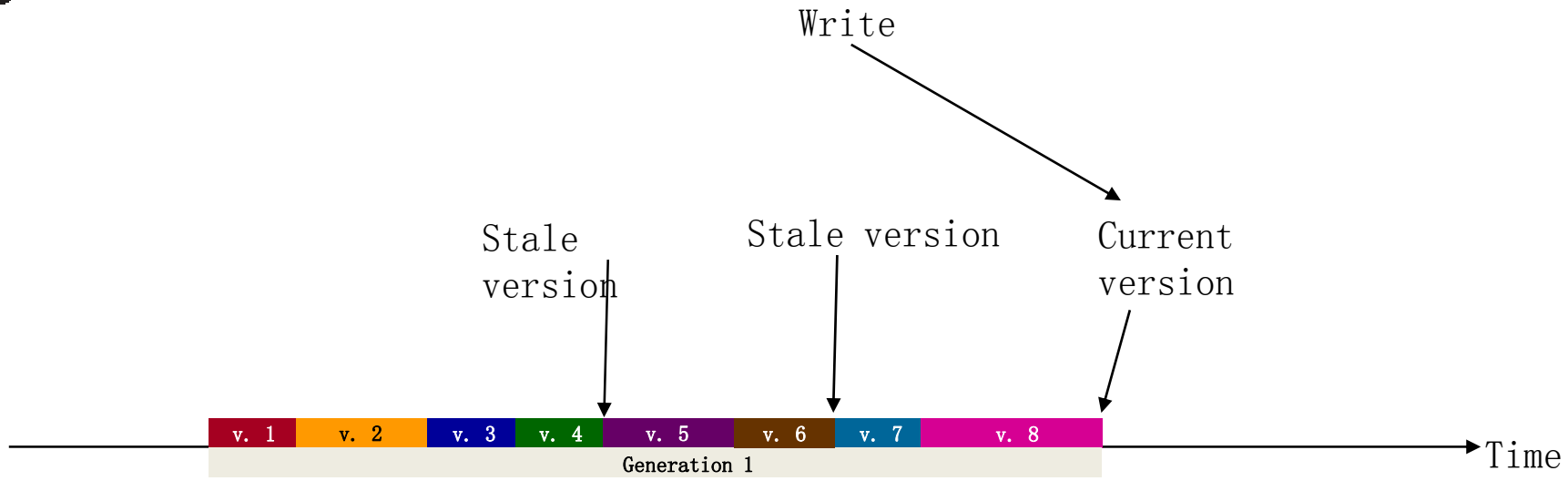
User	Status
Alice	???

East

Record Timeline

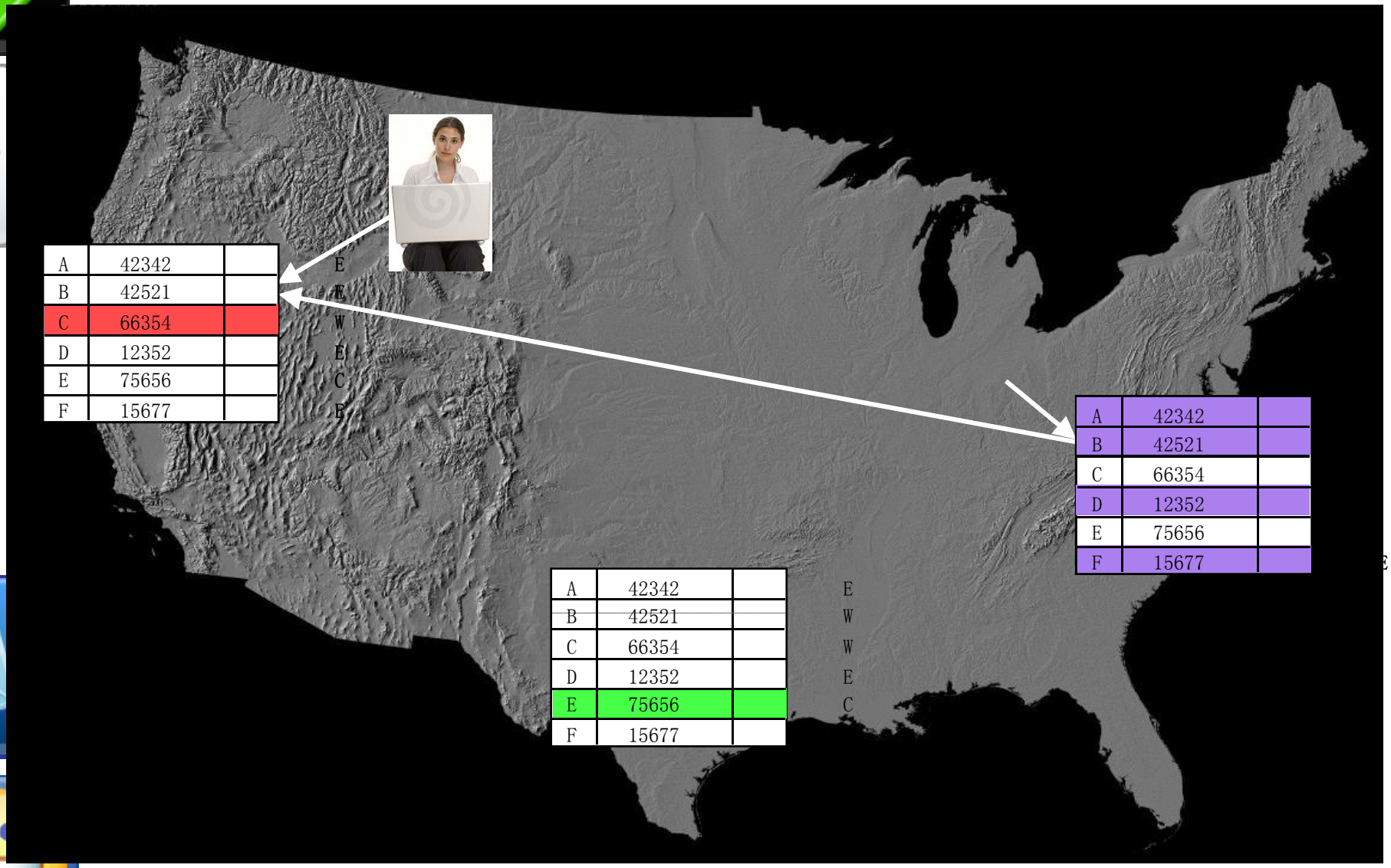


PNUTS Consistency Model



Achieved via per-record primary copy protocol
(To maximize availability, record masterships automatically transferred if site fails)
Can be selectively weakened to eventual consistency
(local writes that are reconciled using version vectors)

Record Master



Consistency Techniques

Per-record mastering

- Each record is assigned a “master region”
  - May differ between records
- Updates to the record forwarded to the master region
- Ensures consistent ordering of updates

Tablet-level mastering

- Each tablet is assigned a “master region”
- Inserts and deletes of records forwarded to the master region
- Master region decides tablet splits

These details are hidden from the application

- Except for the latency impact!

Consistency Techniques

Consistency

Availability

Primary Key Constraint + Record Timeline

- Each tablet is assigned a “master region”
- Inserts of records forwarded to the master region
- Inserts and updates could fail during outages*

Record Timeline Consistency

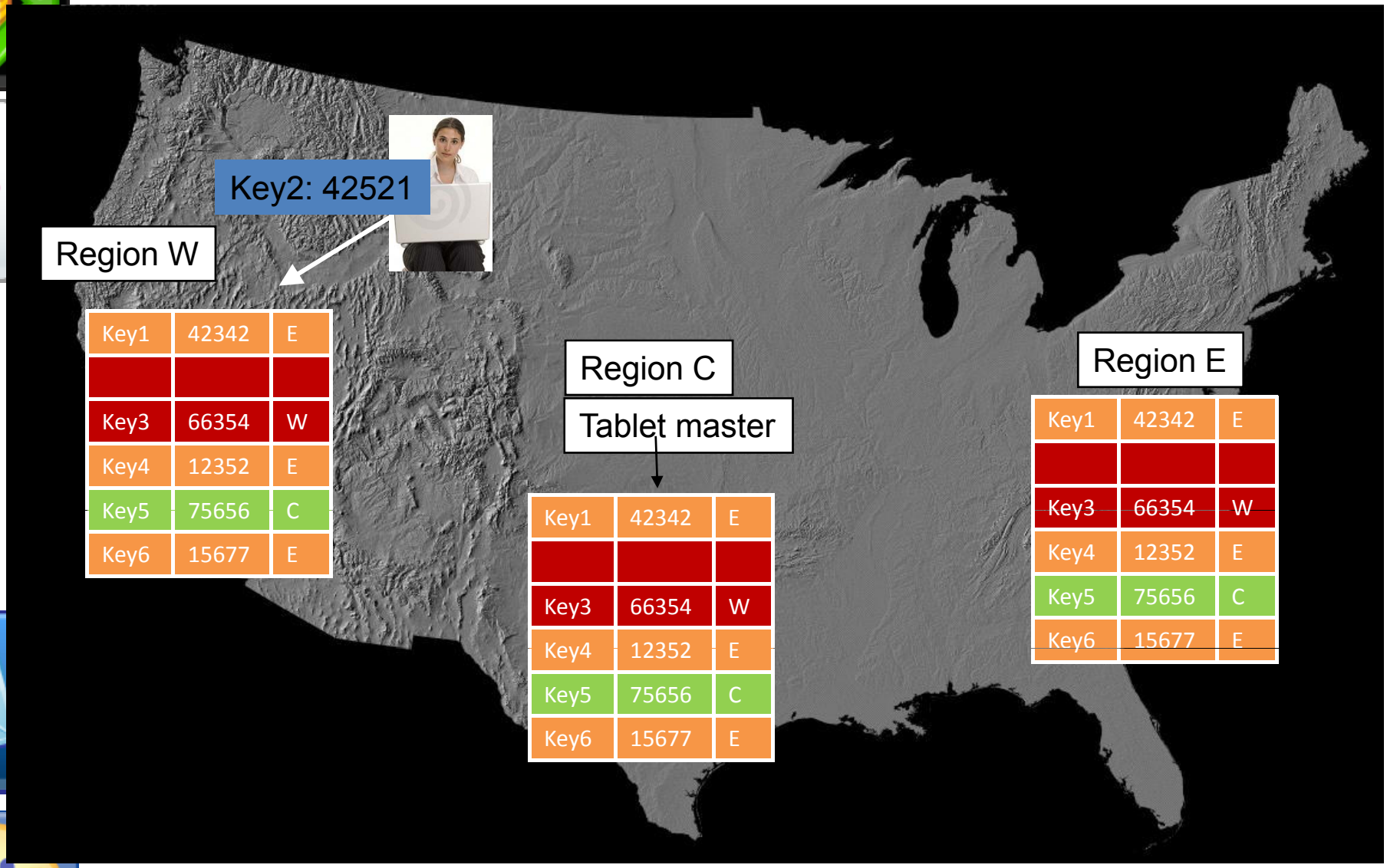
- Each record is assigned a “master region”
- Updates to the record forwarded to the master region
- Inserts succeed, but updates could fail during outages*

Eventual Consistency

- Low latency updates and inserts done locally
- Per field timestamp used to merge updates

★ In case of SU or data center failure. We have failover tools!
★ Reads always will be sent to another region

Tablet Master



Tablet Mastership

Tablet master

Region W

Key1	42342	E
Key3	66354	W
Key4	12352	E
Key5	75656	C
Key6	15677	E

Step 1: Forward
Req to Tablet Master

Region C

Key1	42342	E
Key3	66354	W
Key4	12352	E
Key5	75656	C
Key6	15677	E

Region E

Key1	42342	E
Key3	66354	W
Key4	12352	E
Key5	75656	C
Key6	15677	E

Step 2: Apply
Insert to Tablet Master

Key1	42342	E
Key2	42521	W
~~Key3~~	~~66354~~	~~W~~
Key4	12352	E
Key5	75656	C
Key6	15677	E

Step 3: Replicate
Insert to Other Sites

Step 4: Apply
Insert at Rec Master

Key1	42342	E
Key2	42521	W
Key3	66354	W
Key4	12352	E
Key5	75656	C
Key6	15677	E

Key1	42342	E
Key2	42521	W
Key3	66354	W
Key4	12352	E
Key5	75656	C
Key6	15677	E

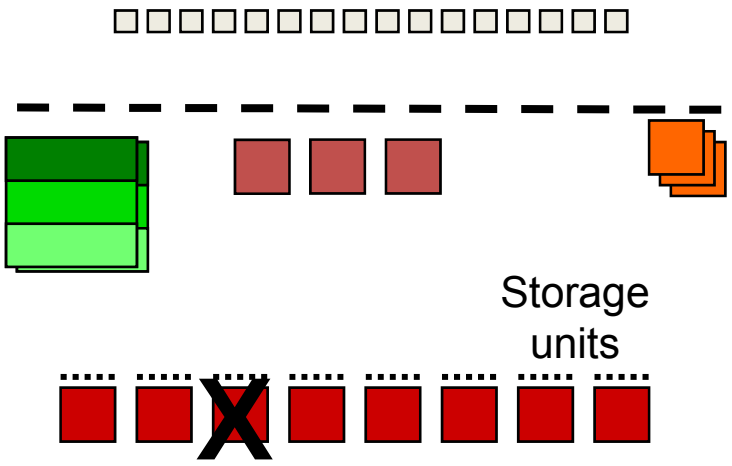
Generalizing Record Timelines to Partition Timelines

Record ➡ Partitition of records with same key

- Tablet splits must respect partition boundaries
- Intra-partition ACID transactions can be done easily now
  - Single machine transactions!
  - With composite keys, this captures Azure and Google AE models
- Each partition is assigned a “master region”
  - May differ between partitions
- Updates to the partition forwarded to the master region
- Ensures consistent ordering of updates across nodes

AVAILABILITY

Possible Failure Modes



Failure type

Storage unit

Consistency impact

None

Availability impact

Degraded service (forwards) for some data.
Updates and inserts fail for some records

Resolution

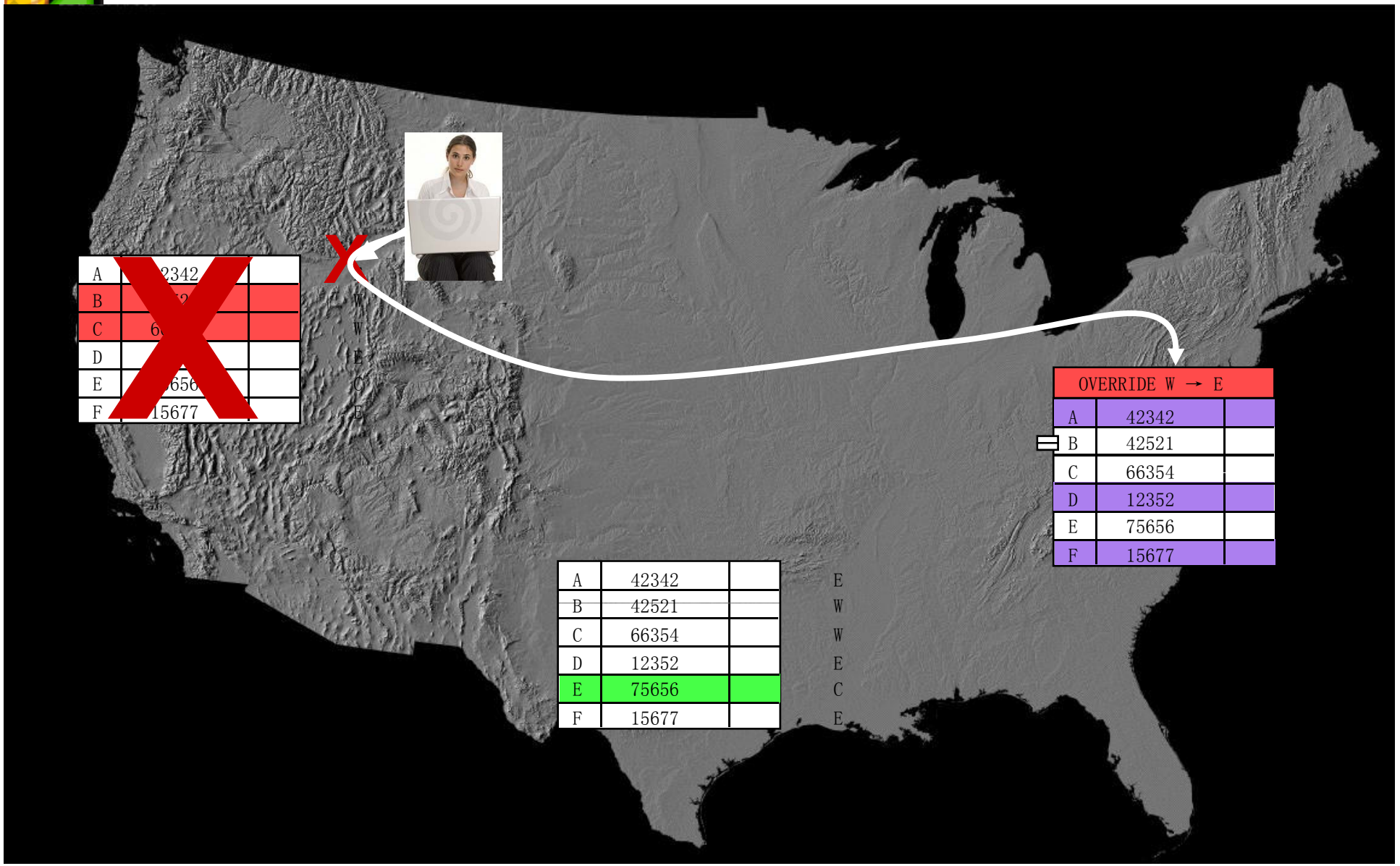
If data not lost: Reboot machine

If data lost: Copy lost tablets from a remote replica

Time to resolve

If data lost, hours or less (depending on tablet size and colo location). If no data lost, minutes.

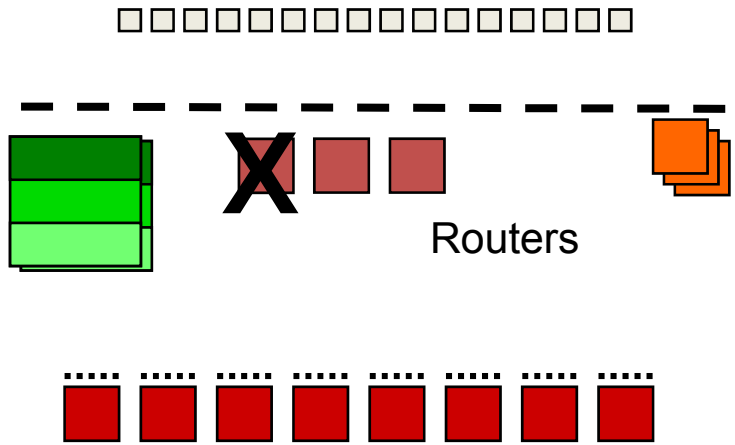
Coping With Failures



Possible Failure Modes

Failure type

Router



Consistency impact

None

Availability impact

None

Resolution

Boot router

Time to resolve

Minutes

Possible Failure Modes

Failure

Tablet controller

Consistency impact

None

Availability impact

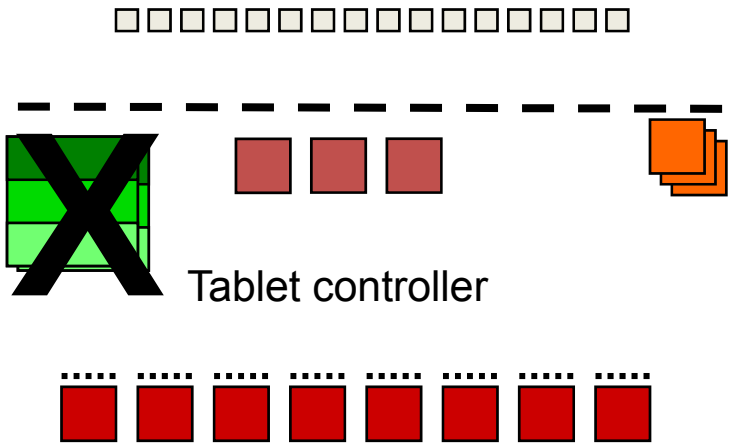
Some actions (e.g., tablet copy) will be blocked

Resolution

Start secondary controller

Time to resolve

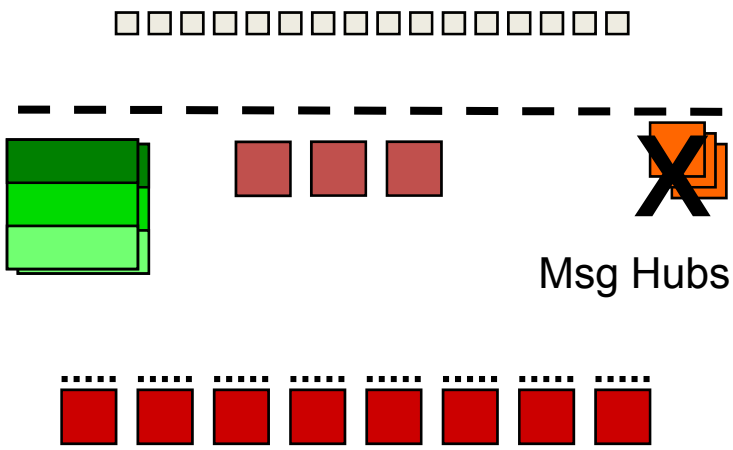
Minutes



Possible Failure Modes

Failure

One msg hub node



Consistency impact

None

Availability impact

Writes fail for some records until a new secondary node takes over

Resolution

Create new primary or secondary for lost topics

Time to resolve

Minutes

Possible Failure Modes

Failure

Colo power outage or partition

Consistency impact

Option to allow “relaxed consistency” to improve availability

Availability impact

Some inserts, updates and deletes cannot succeed

Some critical reads fail

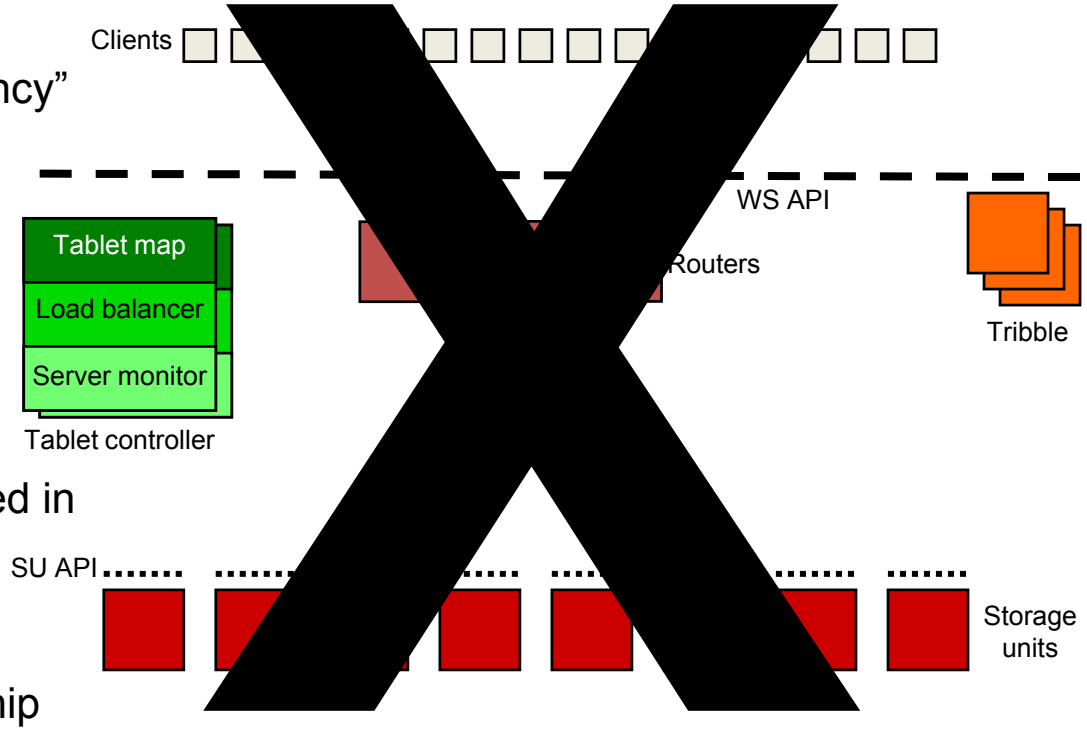
Option to allow updates to proceed in “relaxed consistency mode”

Resolution

Major overrides to force mastership transfer; discard conflicting updates

Time to resolve

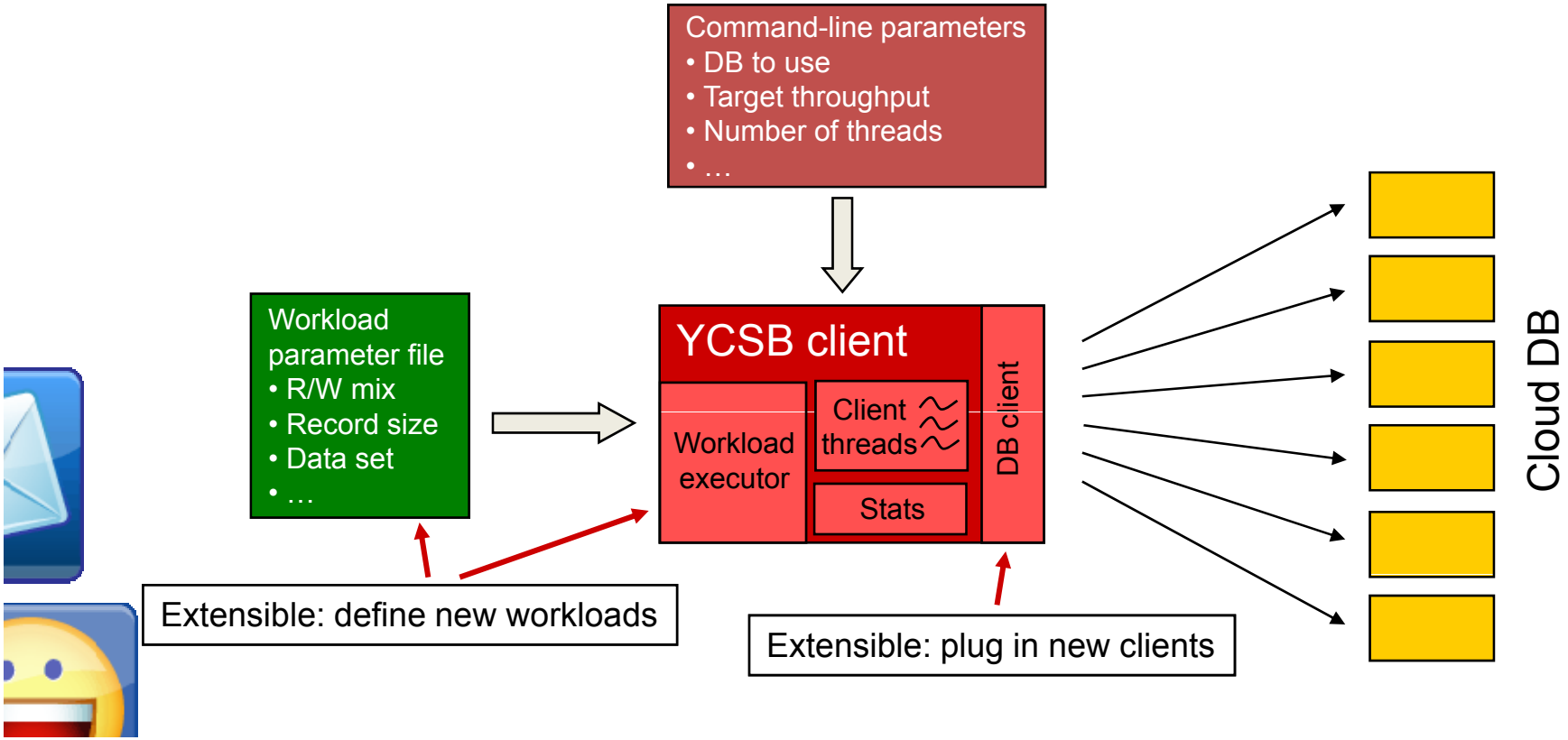
Hours



YCS Benchmark Tool

Java application

- Many systems have Java APIs
- Other systems via HTTP/REST, JNI or some other solution



Further Reading on PNUTS

Efficient Bulk Insertion into a Distributed Ordered Table (SIGMOD 2008)
Adam Silberstein, Brian Cooper, Utkarsh Srivastava, Erik Vee,
Ramana Yerneni, Raghu Ramakrishnan

PNUTS: Yahoo!'s Hosted Data Serving Platform (VLDB 2008)
Brian Cooper, Raghu Ramakrishnan, Utkarsh Srivastava,
Adam Silberstein, Phil Bohannon, Hans-Arno Jacobsen,
Nick Puz, Daniel Weaver, Ramana Yerneni

Asynchronous View Maintenance for VLSD Databases (SIGMOD 2009)
Parag Agrawal, Adam Silberstein, Brian F. Cooper, Utkarsh Srivastava and
Raghu Ramakrishnan

Cloud Storage Design in a PNUTShell **(Beautiful Data, O'Reilly Media, 2009)**
Brian F. Cooper, Raghu Ramakrishnan, and Utkarsh Srivastava

Adaptively Parallelizing Distributed Range Queries (VLDB 2009)
Ymir Vigfusson, Adam Silberstein, Brian Cooper, Rodrigo Fonseca

A Batch of PNUTS: Experiences Connecting Cloud Batch and Serving Systems (SIGMOD 2011)
Adam Silberstein et al.

Summary

- Yahoo! has an extensive cloud computing environment
  - Major open source contributor
  - Currently, focused on internal developers
  - Foundation layer for all Yahoo! products, e.g., Mail, Front Page, Search, Groups …