Big Data in Real-Time *at Twitter*

by Nick Kallen (@nk)

Follow along

http://www.slideshare.net/nkallen/qcon

What is Real-Time Data?

- On-line queries for a single **web request**
- **Off-line** computations with *very* **low latency**
- Latency and throughput are equally important
- **Not talking about Hadoop** and other high-latency, Big Data tools

The four data problems

- **Tweets**
- Timelines
- Social graphs
- Search indices

Your facial hair tells me you revel in celibacy.

11:33 AM Sep 25th, 2009 via web

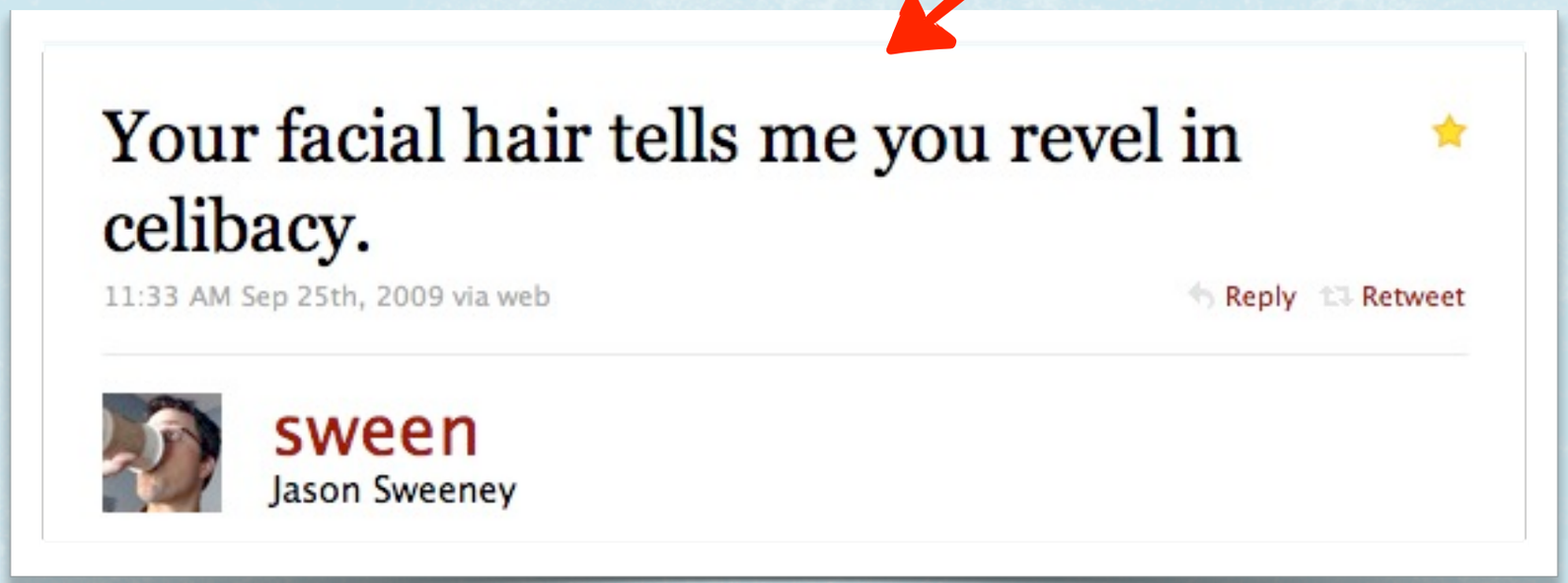
Reply Retweet

sween
Jason Sweeney

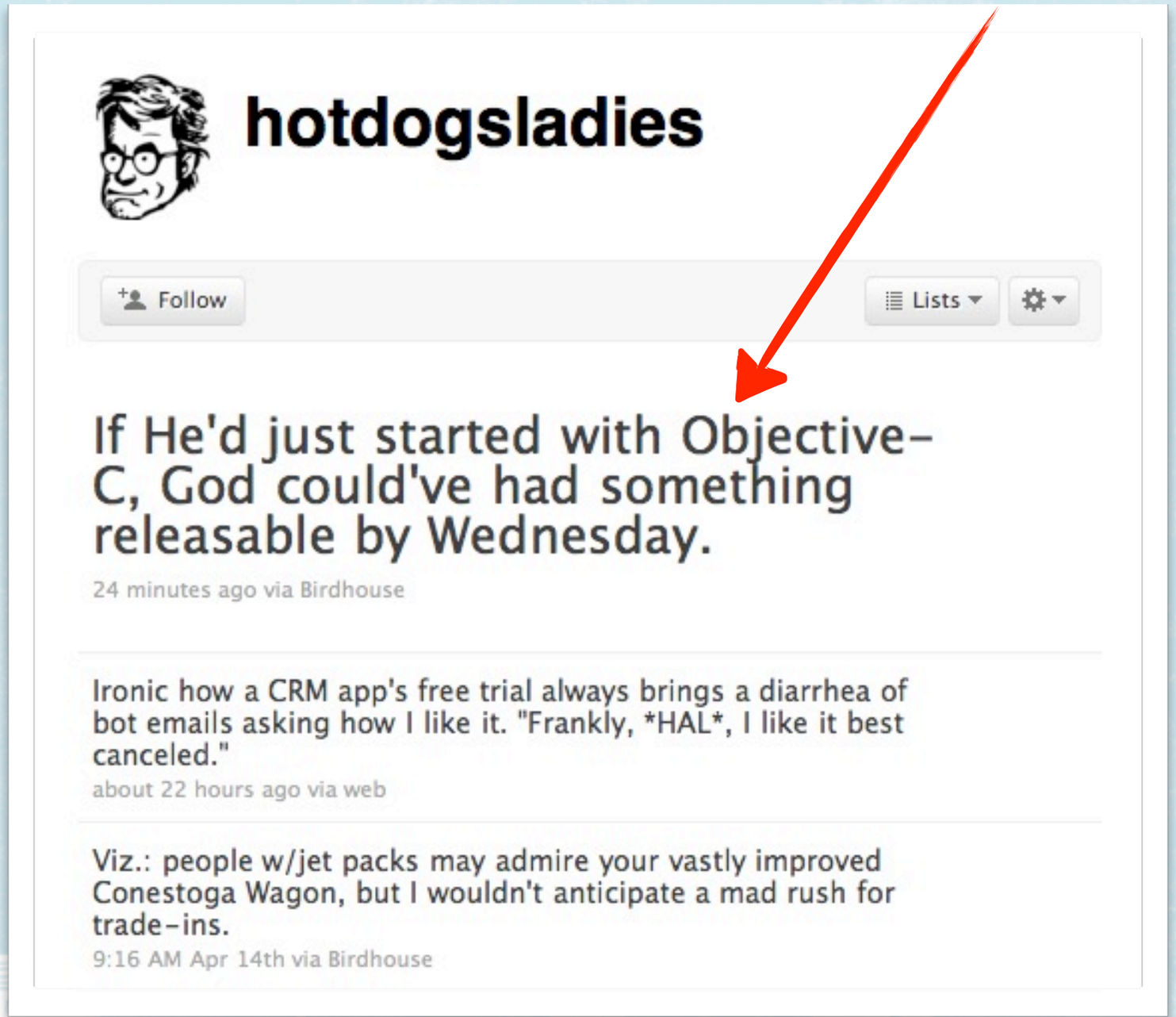
What is a Tweet?

- 140 character message, plus some metadata
- Query patterns:
  - by **id**
  - by **author**
  - (also @replies, but not discussed here)
- Row Storage

Find by primary key: 4376167936



Find all by user_id: 749863



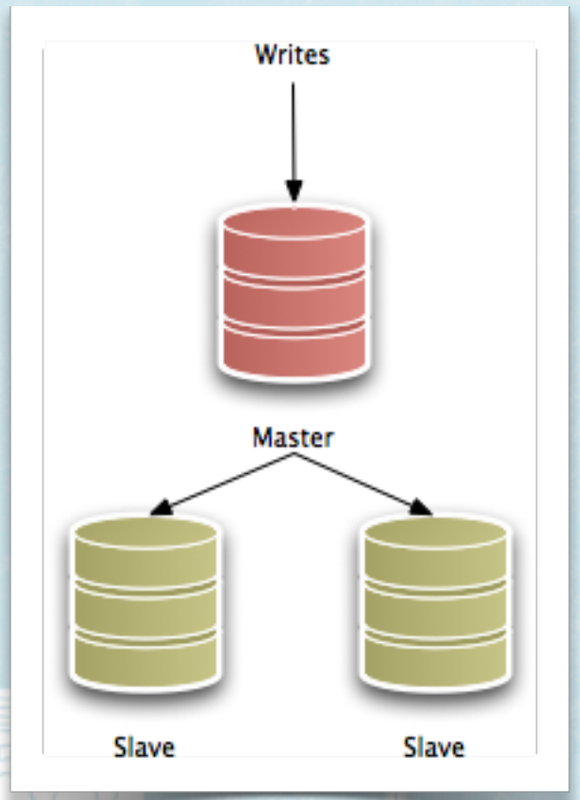
Original Implementation

id	user_id	text	created_at
20	12	just setting up my twttr	2006-03-21 20:50:14
29	12	inviting coworkers	2006-03-21 21:02:56
34	16	Oh shit, I just twittered a little.	2006-03-21 21:08:09

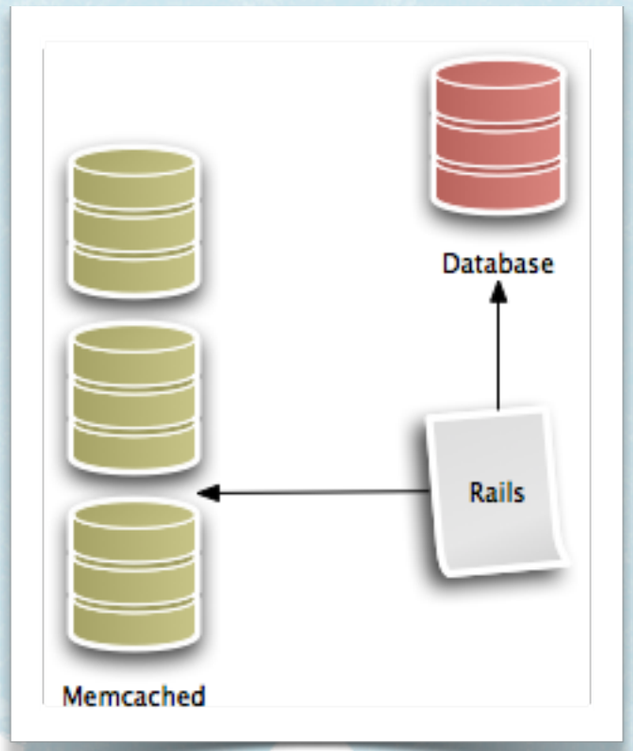
- **Relational**
- **Single table**, vertically scaled
- **Master-Slave** replication and **Memcached** for read throughput.

Original Implementation

Master-Slave Replication



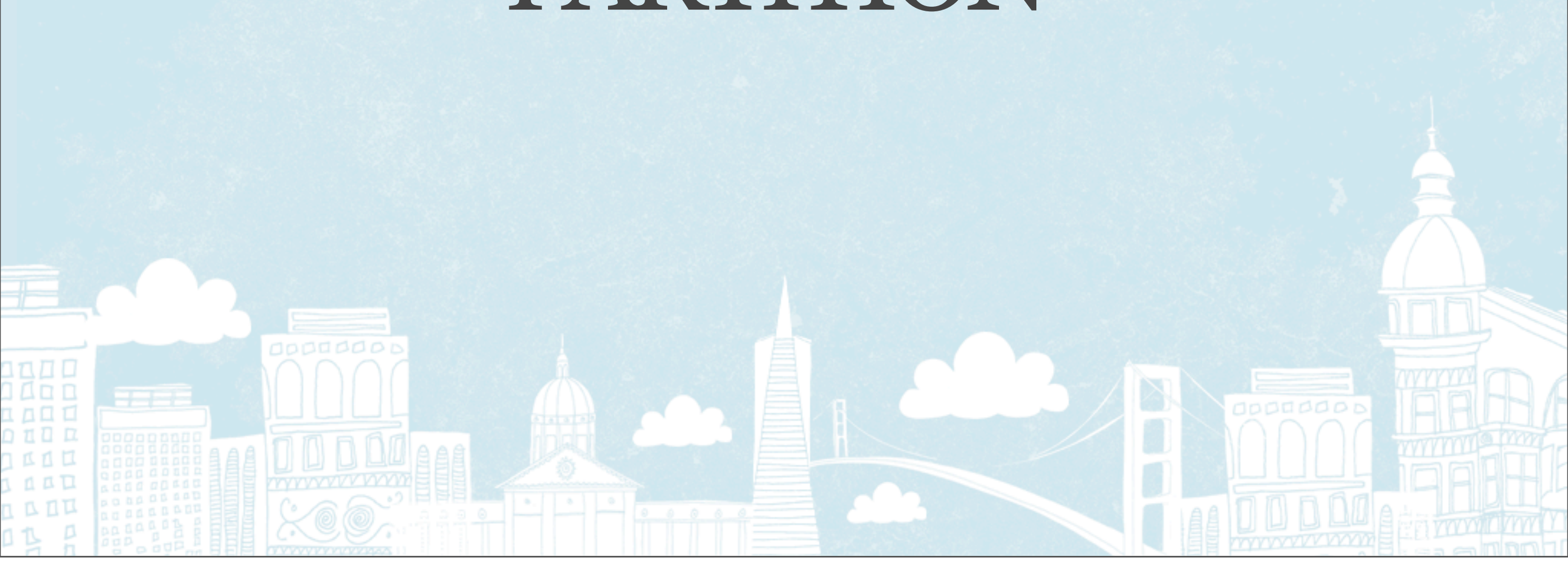
Memcached for reads



Problems w/ solution

- **Disk space**: did not want to support disk arrays larger than 800GB
- At 2,954,291,678 tweets, disk was over 90% utilized.

PARTITION



Possible implementations

Partition by primary key

Partition 1

id	user_id
20	...
22	...
24	...

Partition 2

id	user_id
21	...
23	...
25	...

Possible implementations

Partition by primary key

Partition 1

id	user_id
20	...
22	...
24	...

Partition 2

id	user_id
21	...
23	...
25	...

Finding recent tweets by user_id queries N partitions

Possible implementations

Partition by user id

Partition 1

id	user_id
...	1
...	1
...	3

Partition 2

id	user_id
21	2
23	2
25	2

Possible implementations

Partition by user id

Partition 1

id	user_id
...	1
...	1
...	3

Partition 2

id	user_id
21	2
23	2
25	2

Finding a tweet by id queries N partitions

Current Implementation

Partition by time

Partition 2

id	user_id
24	...
23	...

Partition 1

id	user_id
22	...
21	...

Current Implementation

Partition by time

Partition 2

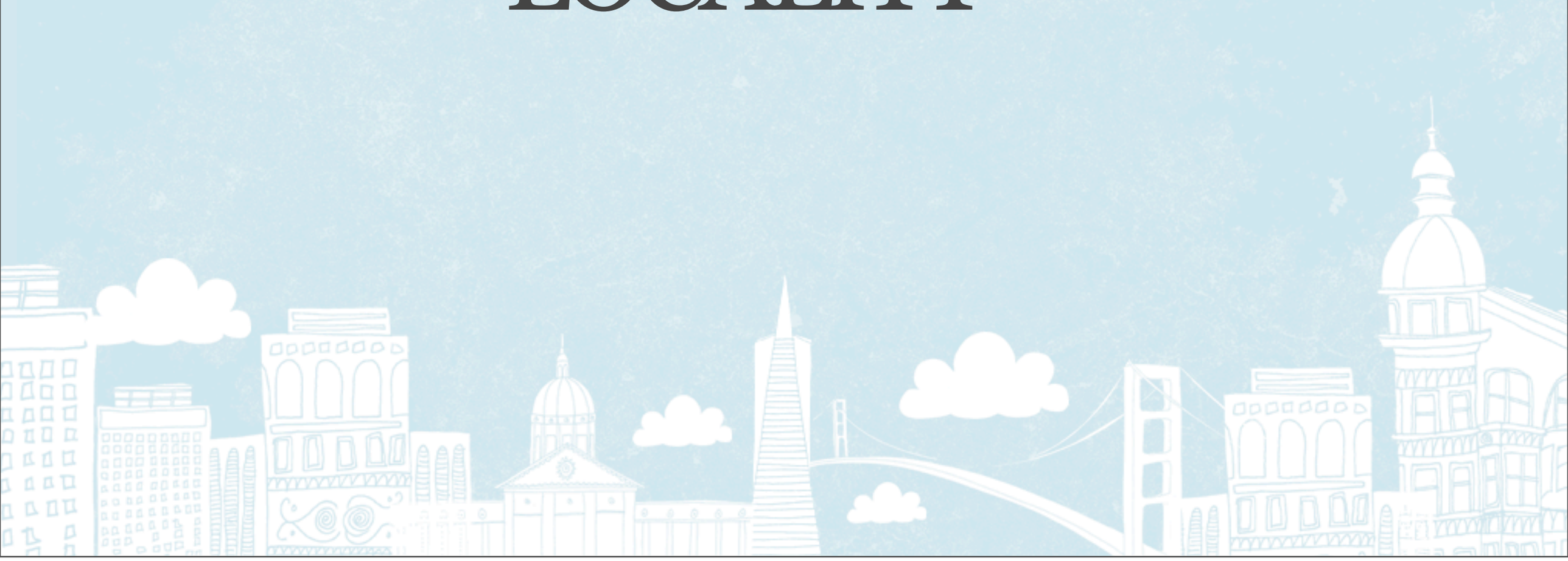
id	user_id
24	...
23	...

Partition 1

id	user_id
22	...
21	...

Queries try each partition in order until enough data is accumulated

LOCALITY



Low Latency

	PK Lookup
Memcached	1ms
MySQL	<10ms*

* Depends on the number of partitions searched

Principles

- Partition and index
- Exploit **locality** (in this case, **temporal** locality)
  - New tweets are requested most frequently, so usually only 1 partition is checked

Problems w/ solution

- Write throughput
- Have encountered **deadlocks** in MySQL at crazy tweet velocity
- Creating a new temporal shard is a manual process and takes too long; it involves setting up a parallel replication hierarchy. Our DBA hates us

Future solution

Partition k1

id	user_id
20	...
22	...

Partition k2

id	user_id
21	...
23	...

Partition u1

user_id	ids
12	20, 21, ...
14	25, 32, ...

Partition u2

user_id	ids
13	48, 27, ...
15	23, 51, ...

- **Cassandra** (non-relational)
- Primary Key partitioning
- Manual **secondary index** on user_id
- Memcached for 90+% of reads

The four data problems

- ~~Tweets~~
- **Timelines**
- Social graphs
- Search indices

What's happening? 140

Latest: Man I need a Tecate 2 minutes ago

Tweet

Home

missionhipster Man I need a Tecate
2 minutes ago via web

NeonIndian Well even if it was an LP after all.. still stoked on that iphone love.
1:54 PM Apr 14th via web

NeonIndian psychic chasms EP?
12:14 PM Apr 14th via mobile web

NeonIndian Overheard band name of the night: Demon Semen.
11:40 PM Apr 13th via mobile web

NeonIndian kareoke to the max! time to take vengance with some jessie's girl...
10:39 PM Apr 13th via mobile web

Sightglass No sweat, big guy--really. RT @therealjoshcook: sorry about your siphon @sighglass :(
6:49 PM Apr 12th via Birdfeed Reply Retweet

What is a Timeline?

- Sequence of tweet ids
- Query pattern: get by user_id
- Operations:
  - append
  - merge
  - truncate
- High-velocity bounded vector
- Space-based (in-place mutation)

Tweets from 3 different people

What's happening? 140

Latest: Man I need a Tecate 2 minutes ago

Tweet

Home

missionhipster Man I need a Tecate
2 minutes ago via web

NeonIndian Well even if it was an LP after all.. still stoked on that iphone love.
1:54 PM Apr 14th via web

NeonIndian psychic chasms EP?
12:14 PM Apr 14th via mobile web

NeonIndian Overheard band name of the night: Demon Semen.
11:40 PM Apr 13th via mobile web

NeonIndian kareoke to the max! time to take vengance with some jessie's girl...
10:39 PM Apr 13th via mobile web

Sightglass No sweat, big guy--really. RT @therealjoshcook: sorry about your siphon @sighglass :(
6:49 PM Apr 12th via Birdfeed Reply Retweet

Original Implementation

```
SELECT * FROM tweets
WHERE user_id IN
  (SELECT source_id
   FROM followers
   WHERE destination_id = ?)
ORDER BY created_at DESC
LIMIT 20
```

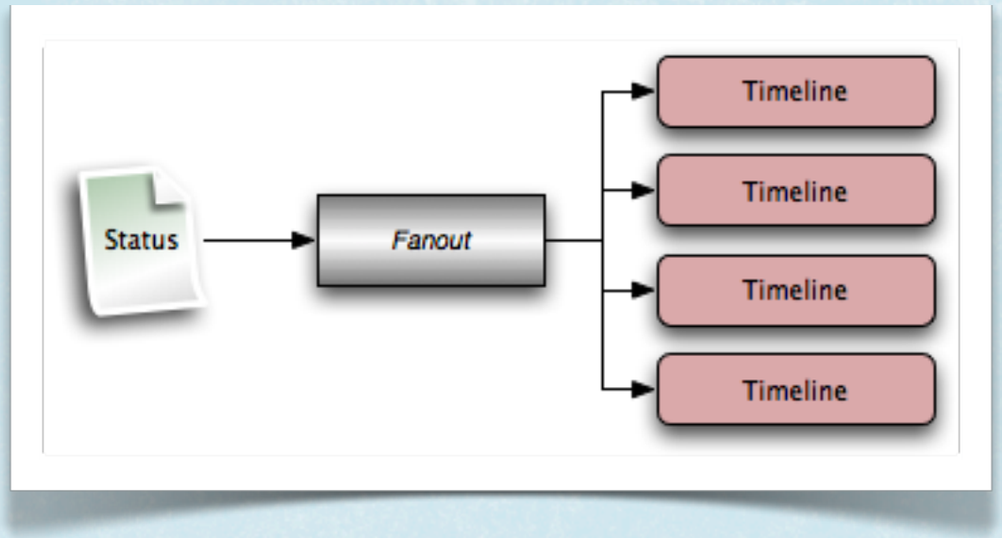
Original Implementation

```
SELECT * FROM tweets
WHERE user_id IN
  (SELECT source_id
   FROM followers
   WHERE destination_id = ?)
ORDER BY created_at DESC
LIMIT 20
```

Crazy slow if you have lots of friends or indices can't be kept in RAM

OFF-LINE *VS.* ONLINE COMPUTATION

Current Implementation



- Sequences stored in **Memcached**
- Fanout off-line, but has a **low latency SLA**
- Truncate at random intervals to ensure bounded length
- On **cache miss**, merge user timelines

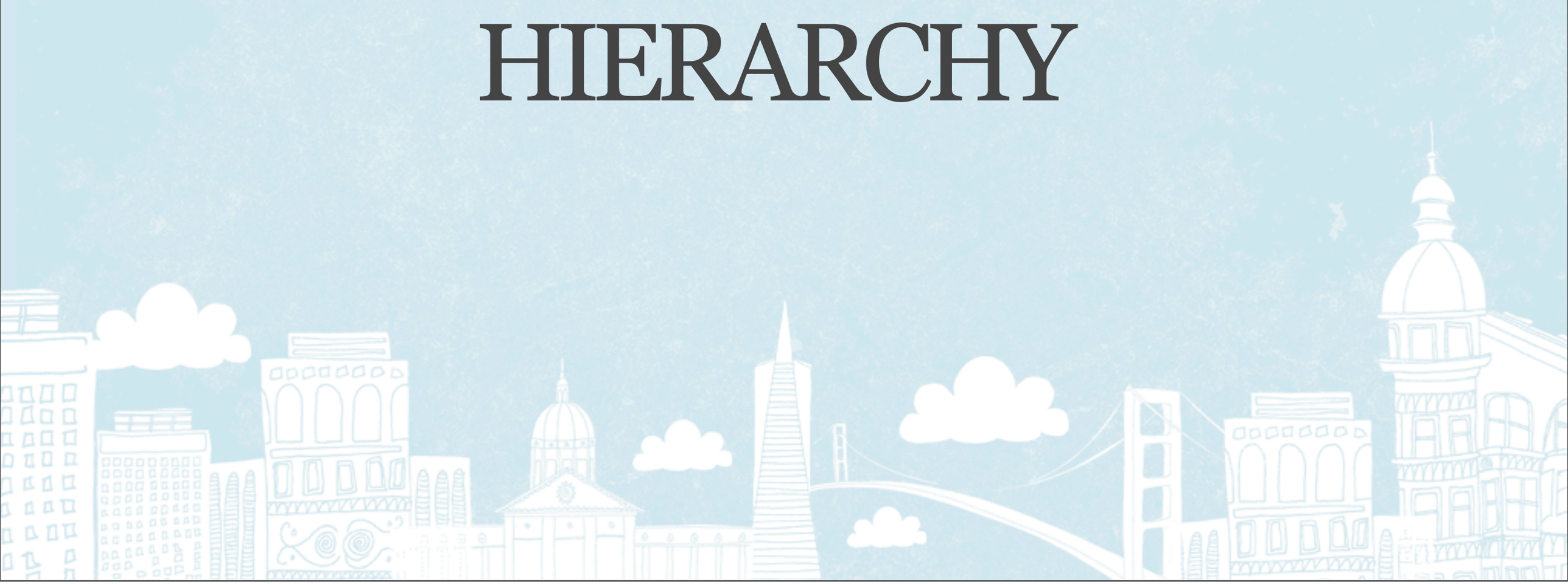
Throughput Statistics

date	average tps	peak tps	fanout ratio	deliveries
10/7/2008	30	120	175:1	21,000
4/15/2010	700	2,000	600:1	1,200,000

1.2m

Deliveries per second

MEMORY HIERARCHY



Possible implementations

- Fanout to **disk**
  - Ridonculous number of IOPS required, even with fancy buffering techniques
  - Cost of rebuilding data from other durable stores not *too* expensive
- Fanout to **memory**
  - Good if cardinality of corpus * bytes/datum not too many GB

Low Latency

get	append	fanout
1ms	1ms	<1s*

* Depends on the number of followers of the tweeter

Principles

- Off-line vs. Online computation
  - The answer to some problems can be **pre-computed** if the amount of work is **bounded** and the query pattern is very limited
- Keep the memory hierarchy in mind
- The efficiency of a system includes the cost of generating data from another source (such as a backup) times the probability of needing to

The four data problems

- ~~Tweets~~
- ~~Timelines~~
- **Social graphs**
- Search indices

Your 15 followers

User / Name

staykreative ✔ Following
Stay Kreative
I'm at Stay Kreative (164 Stanton St, Clinton St, New York). http://4sq.com/ajRPQ1
about 6 hours ago

thebowlingrobot
Alison Dale | 415, 504,505
I subscribed to jamiliya13's channel on YouTube http://www.youtube.com/user/jamiliya13?feature=autoshare_twitter 1:04 PM Apr 14th

immergent ✔ Following
Immergent | Los Angeles
#NameThatSong "Please allow me to introduce myself I'm a man of wealth and taste..." http://ow.ly/1yWA1 about 3 hours ago

thenightatx ✔ Following
The Night | Austin, TX
that was fun thanks y'all :) EP coming out in May, with Radio, Summertime, The Night, Twisted and more http://fb.me/vUx3pyRs 4 days ago

Verified Account

Name ashton kutcher
Location here
Web http://www.facebo...
Bio I make stuff, actually I make up stuff, stories mostly, collaborations of thoughts, dreams, and actions. Thats me.

405	4,764,815	35,887
following	followers	listed

What is a Social Graph?

- List of who follows whom, who blocks whom, etc.
- Operations:
  - Enumerate by time
  - Intersection, Union, Difference
  - Inclusion
  - Cardinality
  - Mass-deletes for spam
- Medium-velocity unbounded vectors
- Complex, predetermined queries

Your 15 followers

User / Name

staykreative ✔ Following
Stay Kreative
I'm at Stay Kreative (164 Stanton St, Clinton St, New York). http://4sq.com/ajRPQ1
about 6 hours ago

thebowlingrobot
Alison Dale | 415, 504,505
I subscribed to jamiliya13's channel on YouTube http://www.youtube.com/user/jamiliya13?feature=autoshare_twitter 1:04 PM Apr 14th

immergent ✔ Following
Immergent | Los Angeles
#NameThatSong "Please allow me to introduce myself I'm a man of wealth and taste..." http://ow.ly/1yWA1 about 3 hours ago

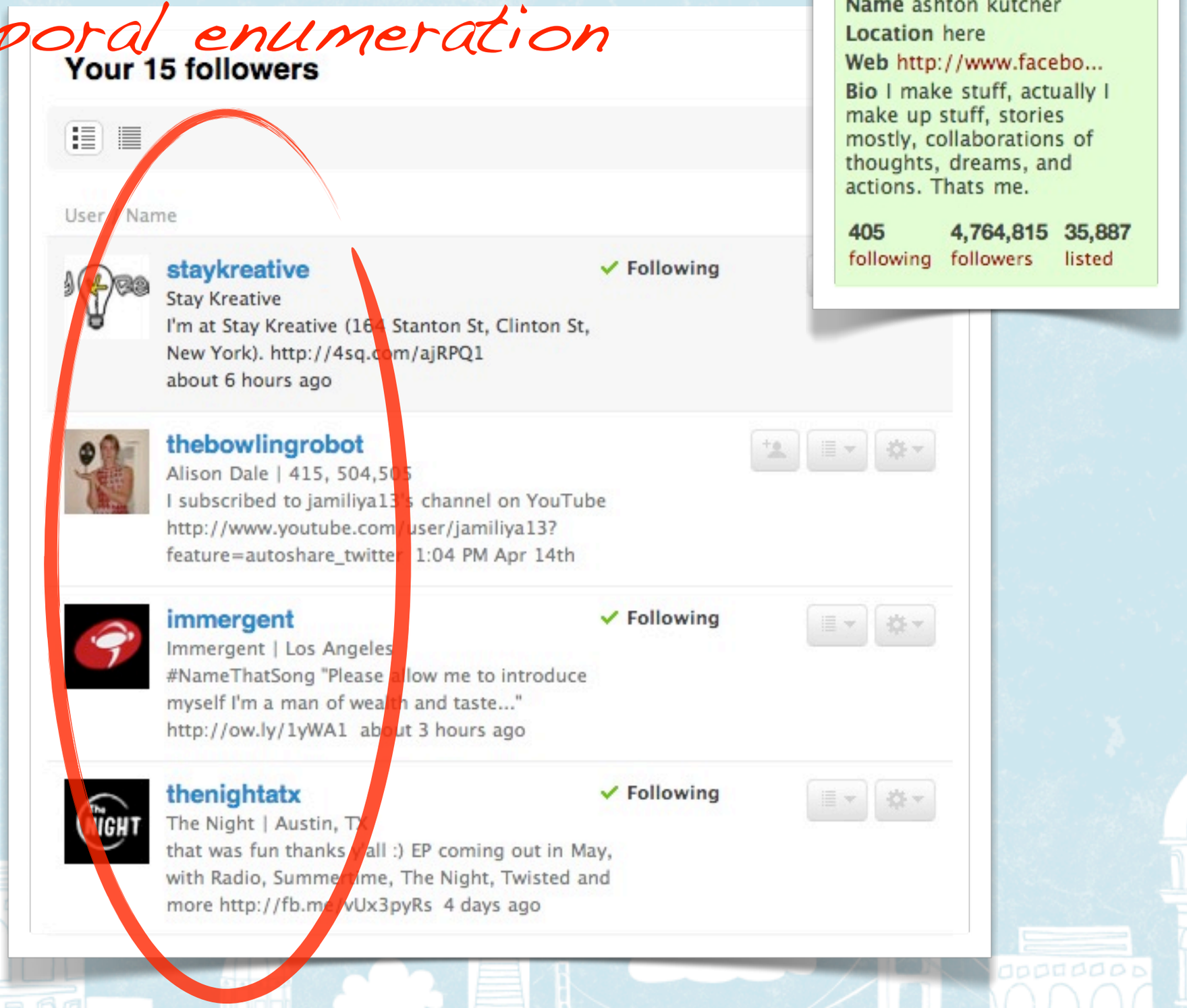
thenightatx ✔ Following
The Night | Austin, TX
that was fun thanks y'all :) EP coming out in May, with Radio, Summertime, The Night, Twisted and more http://fb.me/vUx3pyRs 4 days ago

Verified Account

Name ashton kutcher
Location here
Web http://www.facebo...
Bio I make stuff, actually I make up stuff, stories mostly, collaborations of thoughts, dreams, and actions. Thats me.

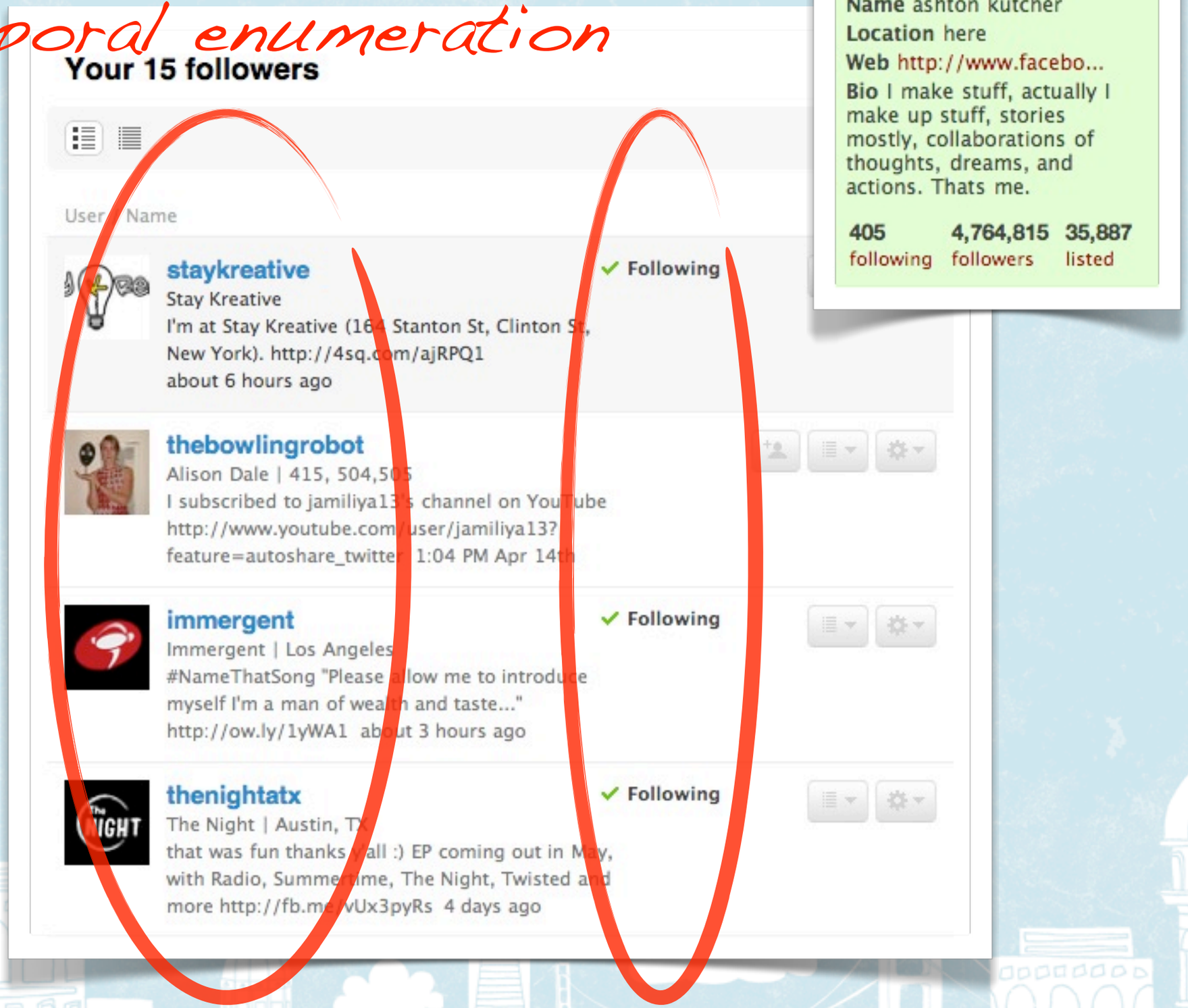
405	4,764,815	35,887
following	followers	listed

Temporal enumeration



Temporal enumeration

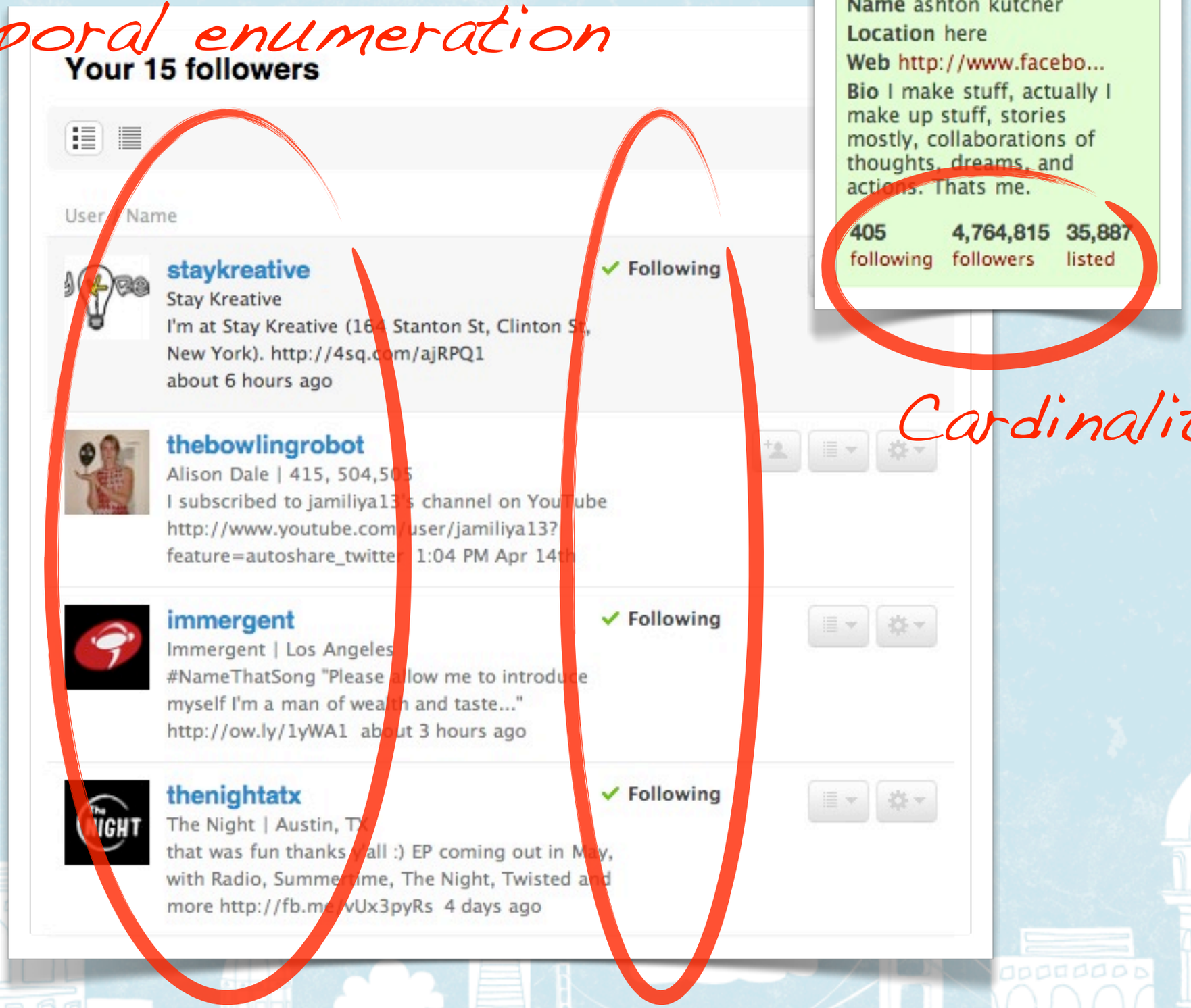
Inclusion



Inclusion

Temporal enumeration

Cardinality

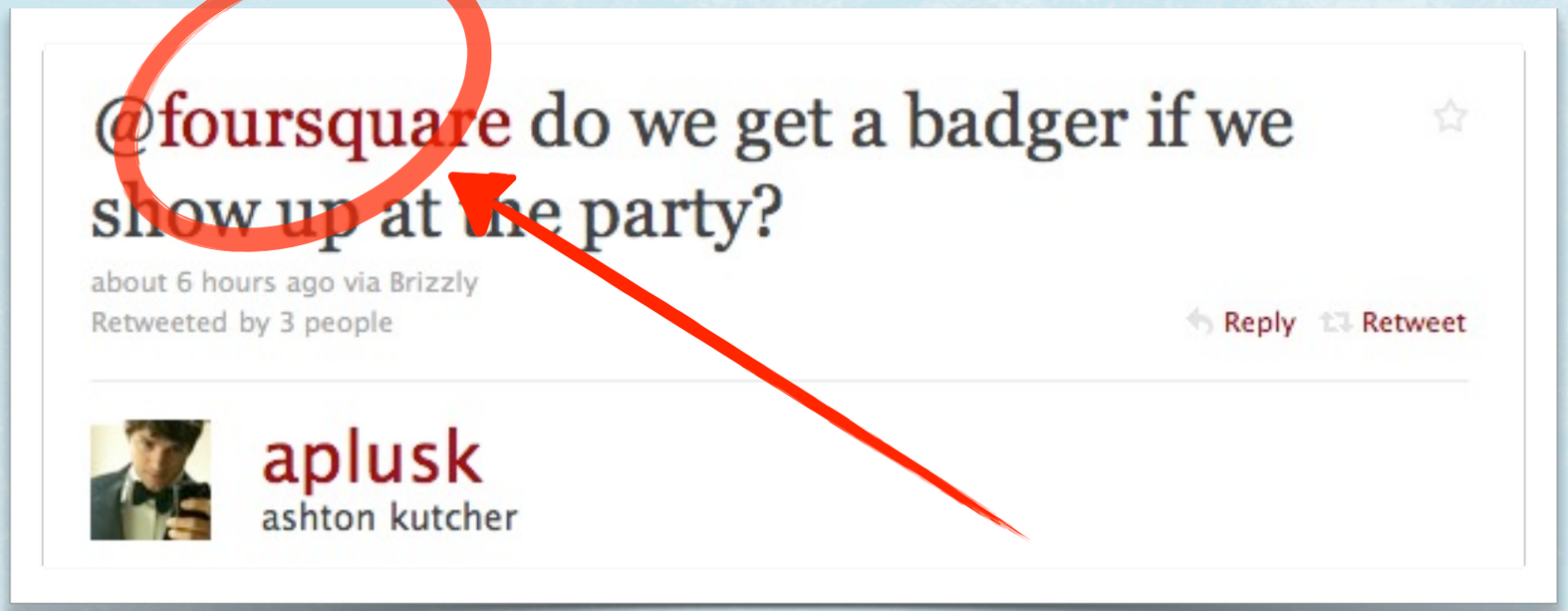


@foursquare do we get a badger if we show up at the party?

about 6 hours ago via Brizzly
Retweeted by 3 people

Reply Retweet

aplusk
ashton kutcher



Intersection: Deliver to people who follow both @aplusk and @foursquare

Original Implementation

source_id	destination_id
20	12
29	12
34	16

- **Single table**, vertically scaled
- **Master-Slave** replication

Index

Original Implementation

source_id	destination_id
20	12
29	12
34	16

- **Single table**, vertically scaled
- **Master-Slave** replication

Original Implementation

Index

Index

source_id	destination_id
20	12
29	12
34	16

- **Single table**, vertically scaled
- **Master-Slave** replication

Problems w/ solution

- Write throughput
- Indices couldn't be kept in RAM

Current solution

Forward

source_id	destination_id	updated_at	x
20	12	20:50:14	x
20	13	20:51:32	
20	16		

Backward

destination_id	source_id	updated_at	x
12	20	20:50:14	x
12	32	20:51:32	
12	16		

- Partitioned by user id
- Edges stored in “forward” and “backward” directions
- **Indexed** by time
- **Indexed** by element (for **set algebra**)
- Denormalized cardinality

Current solution

Forward

source_id	destination_id	updated_at	x
20	12	20:50:14	x
20	13	20:51:32	
20	16		

Backward

destination_id	source_id	updated_at	x
12	20	20:50:14	x
12	32	20:51:32	
12	16		

- Partitioned by user id
- Edges stored in "forward" and "backward" directions
- **Indexed** by time
- **Indexed** by element (for **set algebra**)
- Denormalized cardinality

Partitioned by user

Edges stored in both directions

Current solution

Forward

source_id	destination_id	updated_at	x
20	12	20:50:14	x
20	13	20:51:32	
20	16		

Backward

destination_id	source_id	updated_at	x
12	20	20:50:14	x
12	32	20:51:32	
12	16		

- Partitioned by user id
- Edges stored in "forward" and "backward" directions
- **Indexed** by time
- **Indexed** by element (for **set algebra**)
- Denormalized cardinality

Partitioned by user

Challenges

- Data consistency in the presence of failures
- Write operations are **idempotent**: retry until success
- **Last-Write Wins** for edges
  - (with an **ordering relation on State** for time conflicts)
- Other **commutative** strategies for mass-writes

Low Latency

cardinality	iteration	write ack	write materialize	inclusion
1ms	100edges/ms*	1ms	16ms	1ms

* 2ms lower bound

Principles

- It is not possible to pre-compute set algebra queries
- **Simple** distributed coordination **techniques work**
- **Partition, replicate, index**. Many efficiency and scalability problems are solved the same way

The four data problems

- ~~Tweets~~
- ~~Timelines~~
- ~~Social graphs~~
- **Search indices**

Real-time results for **mountain dew cheetos** Save this search

red_mittens RT @jeffreymax: I'm making soup. Just kidding. I'm pouring a **Mountain Dew** Code Red into a bag of **Cheetos** Puffs. Close though. Pretty close.

1 minute ago from web Reply Retweet

jeffreymax I'm making soup. Just kidding. I'm pouring a **Mountain Dew** Code Red into a bag of **Cheetos** Puffs. Close though. Pretty close.

4 minutes ago from web

EBennett07 For the most part, I try to be healthy.... but sometimes to get the day going all you need is a little **Mountain Dew** and **Cheetos**.

about 10 hours ago from web

gordondunn Buying selection of my favorite life-giving health foods for coming 30th. Knoppers, **mountain dew**, **cheetos**, reese's, etc. Nyum, nyum, nyum.

1 day ago from Tweetie

showalasomelove I swear I have the diet of a 10yr old. Who else would combine hot **cheetos**, twix, honey buns and **mountain dew**? smh.

1 day ago from Twitter for BlackBerry®

JadeWinters **Cheetos** and **Mountain Dew**. The breakfast of champions.

1 day ago from web

What is a Search Index?

- "Find me all tweets with these words in it..."
- Posting list
- Boolean and/or queries
- Complex, ad hoc queries
- Relevance is recency*

* Note: there is a non-real-time component to search, but it is not discussed here

Real-time results for **mountain dew cheetos** Save this search

red_mittens RT @jeffreymax: I'm making soup. Just kidding. I'm pouring a **Mountain Dew** Code Red into a bag of **Cheetos** Puffs. Close though. Pretty close.
1 minute ago from web Reply Retweet

jeffreymax I'm making soup. Just kidding. I'm pouring a **Mountain Dew** Code Red into a bag of **Cheetos** Puffs. Close though. Pretty close.
4 minutes ago from web

EBennett07 For the most part, I try to be healthy.... but sometimes to get the day going all you need is a little **Mountain Dew** and **Cheetos**.
about 10 hours ago from web

gordondunn Buying selection of my favorite life-giving health foods for coming 30th. Knoppers, **mountain dew**, **cheetos**, reese's, etc. Nyum, nyum, nyum.
1 day ago from Tweetie

showalasomelove I swear I have the diet of a 10yr old. Who else would combine hot **cheetos**, twix, honey buns and **mountain dew**? smh.
1 day ago from Twitter for BlackBerry®

JadeWinters **Cheetos** and **Mountain Dew**. The breakfast of champions.
1 day ago from web

Intersection of three posting lists

Original Implementation

term_id	doc_id
20	12
20	86
34	16

- **Single table**, vertically scaled
- **Master-Slave** replication for read throughput

Problems w/ solution

- Index could not be kept in memory

Current Implementation

Partition 2

term_id	doc_id
24	...
23	...

Partition 1

term_id	doc_id
22	...
21	...

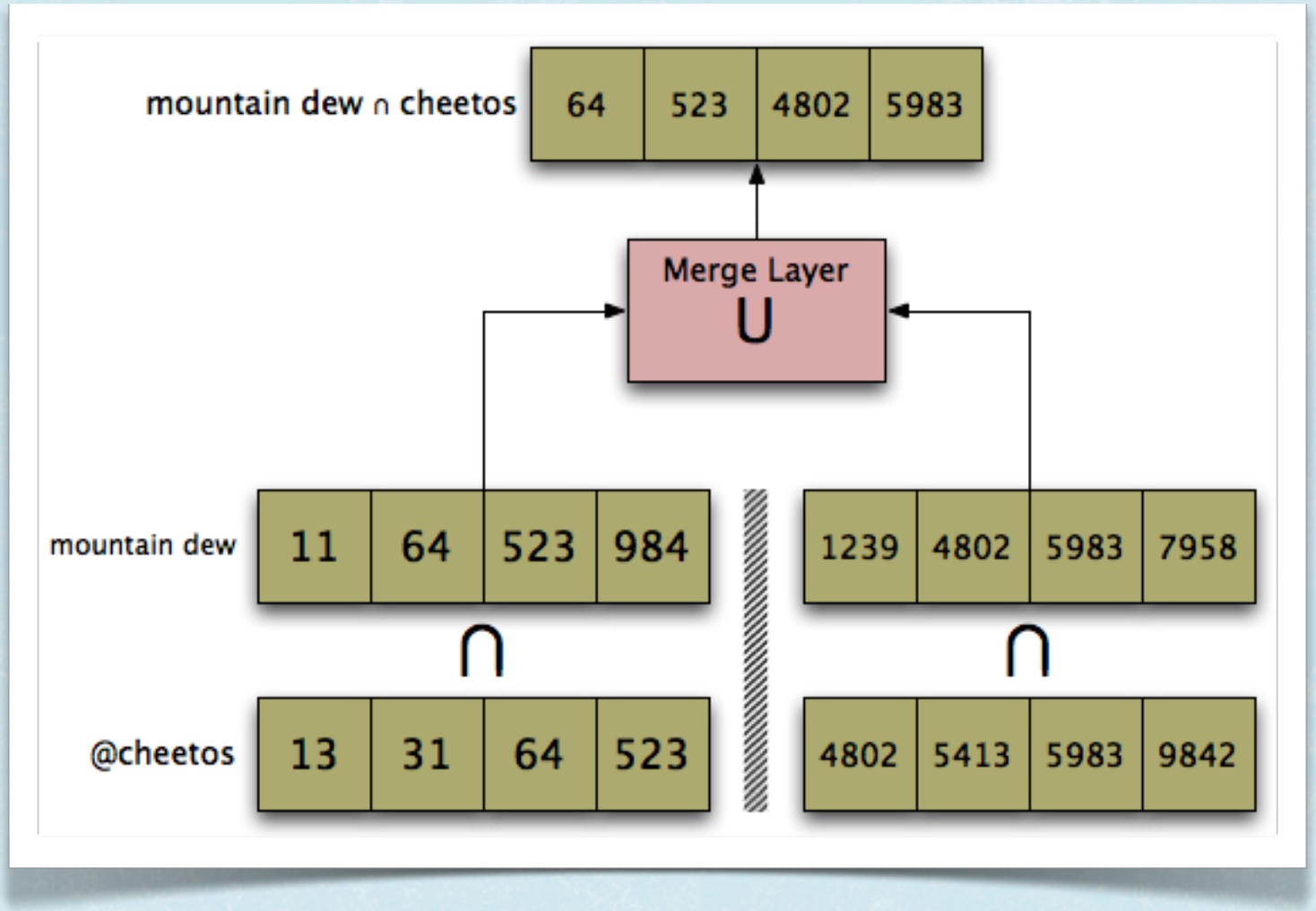
- **Partitioned by time**
- Uses **MySQL**
- Uses delayed key-write

Problems w/ solution

- Write throughput
- Queries for **rare terms** need to search *many* partitions
- Space efficiency/recall
  - MySQL requires lots of memory

DATA NEAR COMPUTATION

Future solution



- Document partitioning
- Time partitioning too
- Merge layer
- **May** use **Lucene** instead of MySQL

Principles

- Partition so that work can be **parallelized**
- Temporal locality is not always enough

The four data problems

- ~~Tweets~~
- ~~Timelines~~
- ~~Social graphs~~
- ~~Search indices~~

Summary Statistics

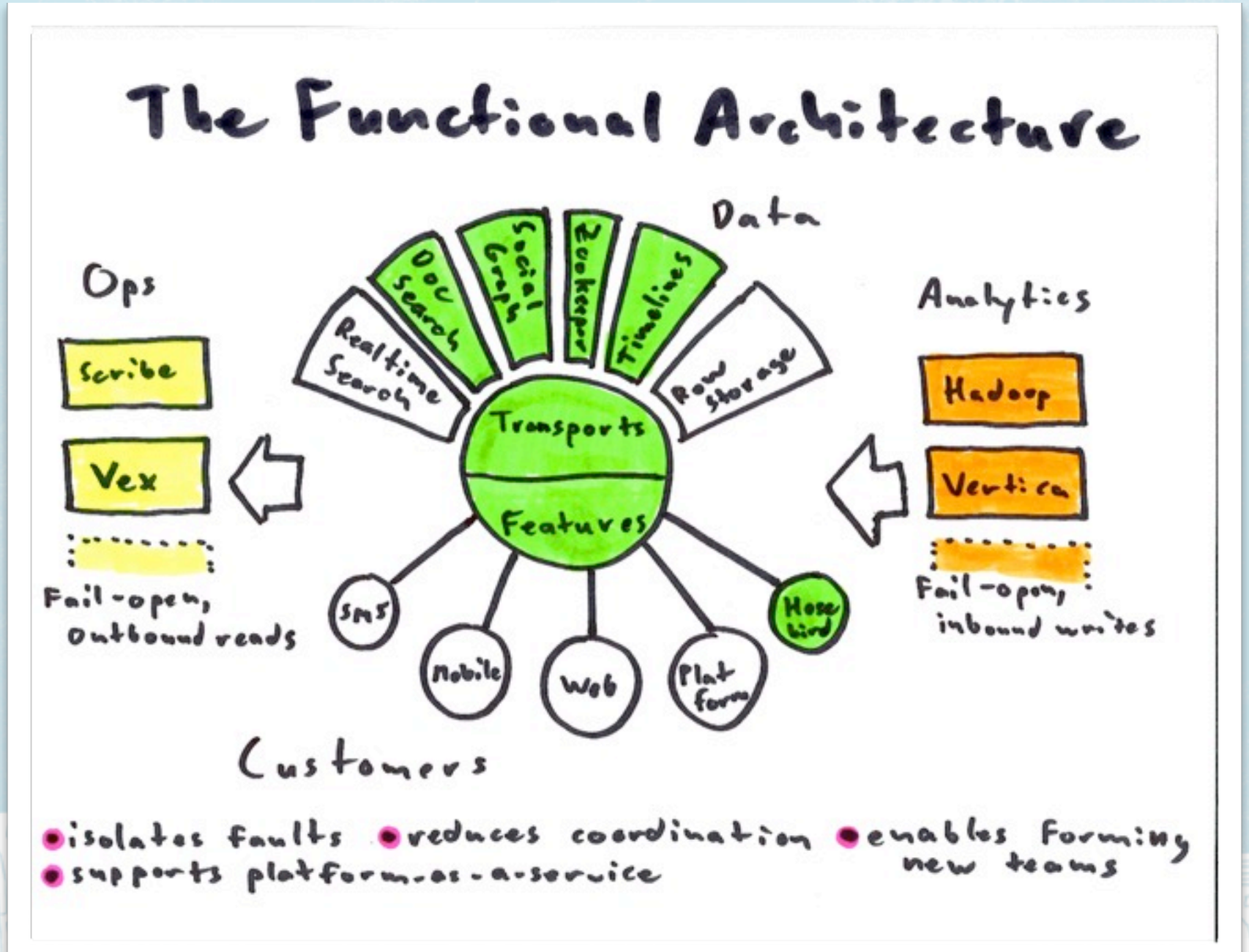
	reads/second	writes/ second	cardinality	bytes/item	durability
Tweets	100k	850	12b	300b	durable
Timelines	80k	1.2m	a lot	3.2k	non
Graphs	100k	20k	13b	110	durable
Search	13k	21k†	315m‡	1k	durable

† tps * 25 postings
‡ 75 partitions * 4.2m tweets

Principles

- All **engineering solutions are transient**
- Nothing's perfect but some solutions are good enough for a while
- Scalability solutions aren't magic. They involve **partitioning, indexing, and replication**
- All data for real-time queries **MUST** be in memory. **Disk is for writes only**.
- Some problems can be solved with **pre-computation**, but a lot can't
- Exploit locality where possible

Appendix


The Functional Architecture
Data
Ops
Analytics
Scribe
Vex
Fail-open,
Outbound reads
Doc search
Social Graph
Zookeeper
Timelines
Realtime Search
Row storage
Transports
Features
SMS
Mobile
Web
Platform
Hose bird
Hadoop
Vertica
Fail-open,
inbound writes
Customers
isolates faults
reduces coordination
enables forming new teams
supports platform-as-a-service