

Navigating the Data Pipeline

Tim Moreton
@timmoreton



In this talk...



The Big Data Pipeline,
and 'Real Time' challenges



A Proposal:
Let's Unify the Pipeline

Big Data pipeline



'Ingest'

'Database'

'Dashboard'

- Mike Driscoll, Metamarkets

**Noisy, low
value events**

**Actionable
observations**

Dashboard

Saved Reports

- Visitors
- Traffic Sources
- Content
- Goals

Settings

Email

Help Resources

- About this Report
- Conversion University
- Common Questions
- Report Finder
- Beta Feedback

Dashboard

Export | Email

Apr 1, 2007 - May 1, 2007

Avg. Time on Site



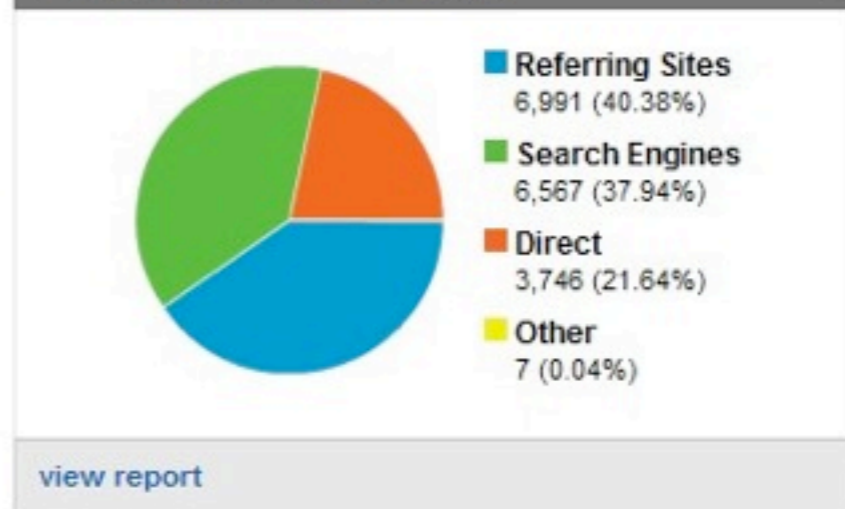
Site Usage



Visitors Overview



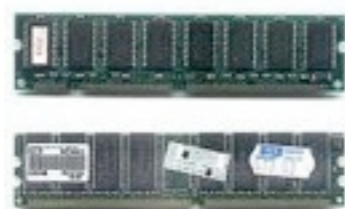
Traffic Sources Overview



'Real Time' Big Data



Value of
timeliness

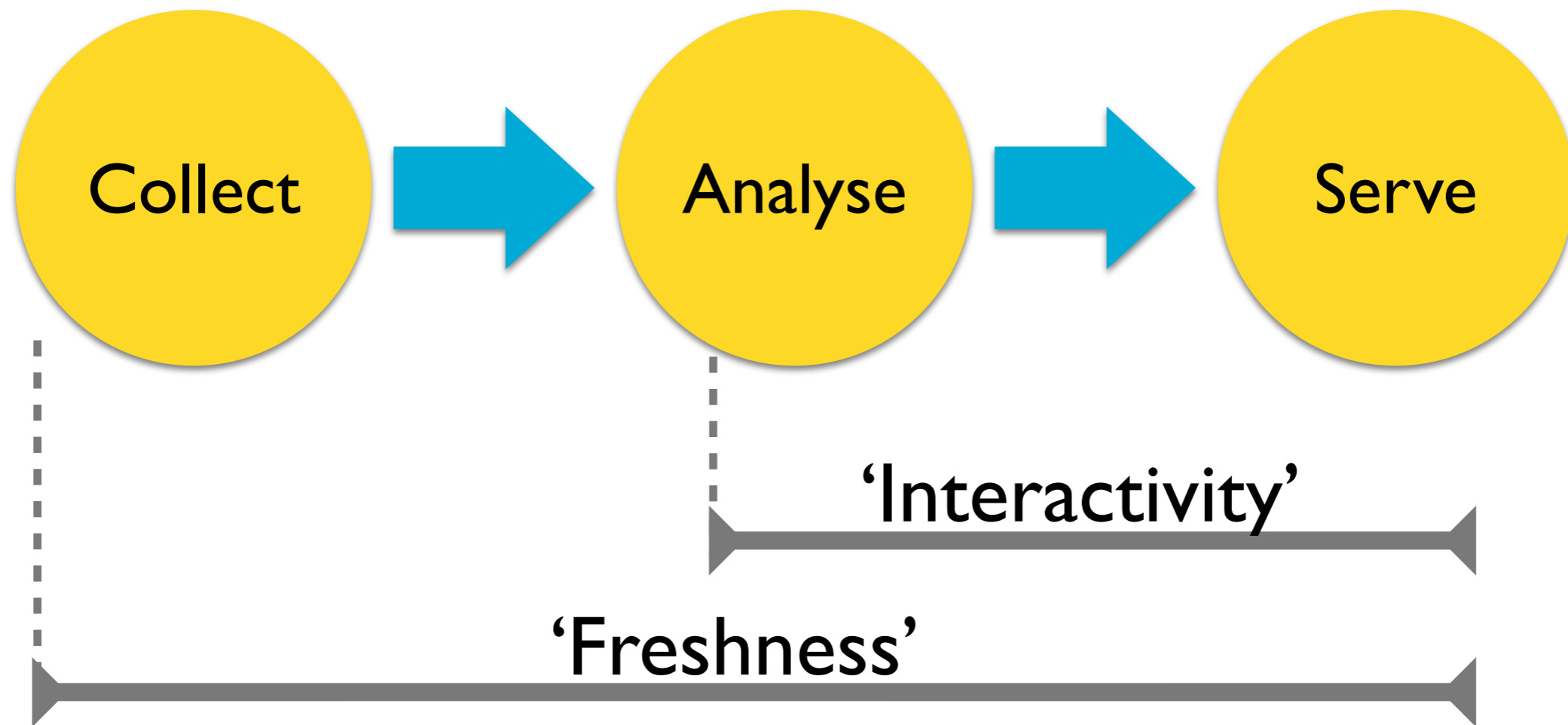


Dataset size



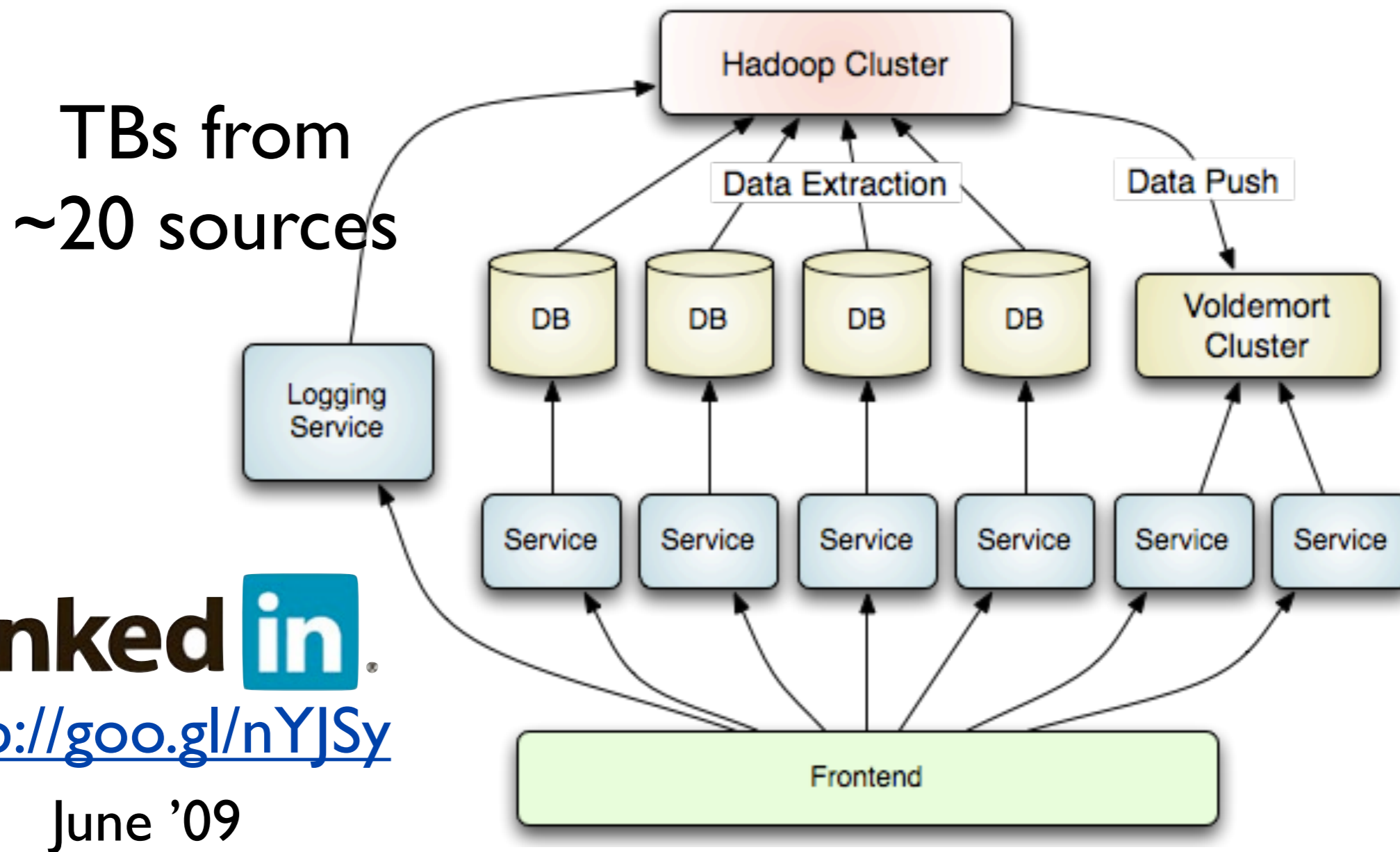
seconds + TBs

'Real Time' pipeline



Challenge: lower latency increases value

Starting point: silos, glue

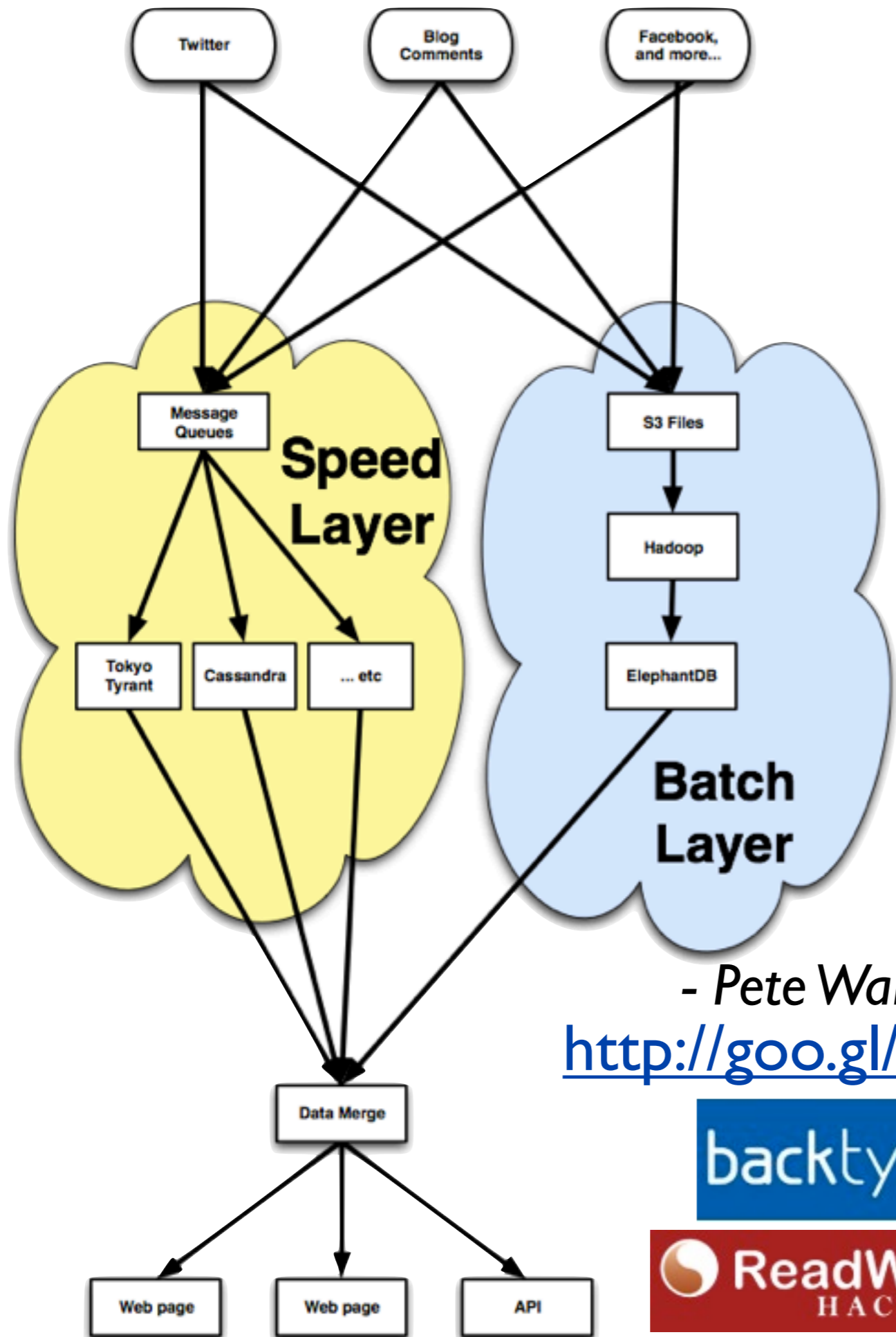


Linked in
<http://goo.gl/nYJSy>

June '09

Or...

approximate results



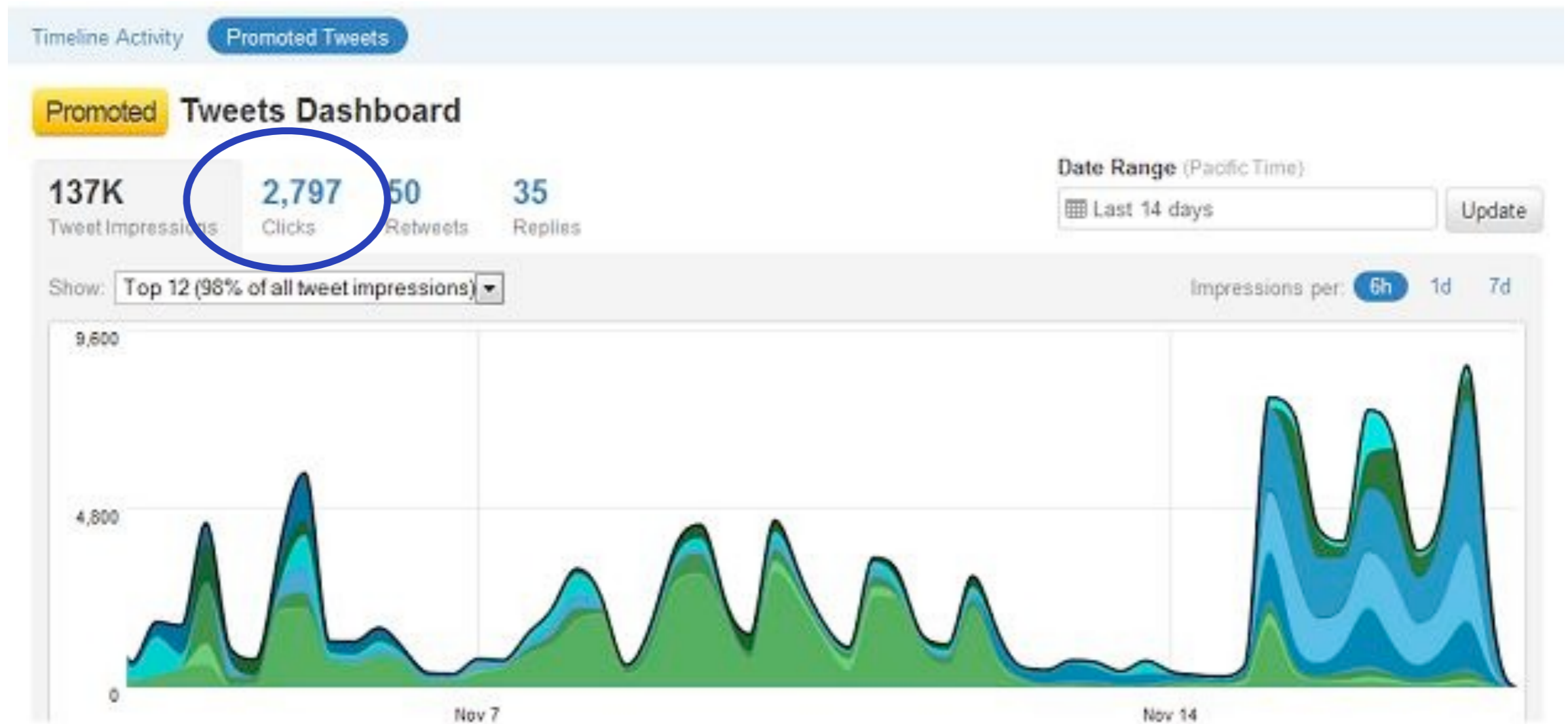
- Isolated pipelines, then merge
- Bound max 'drift'

- Pete Warden

<http://goo.gl/IQVQa>

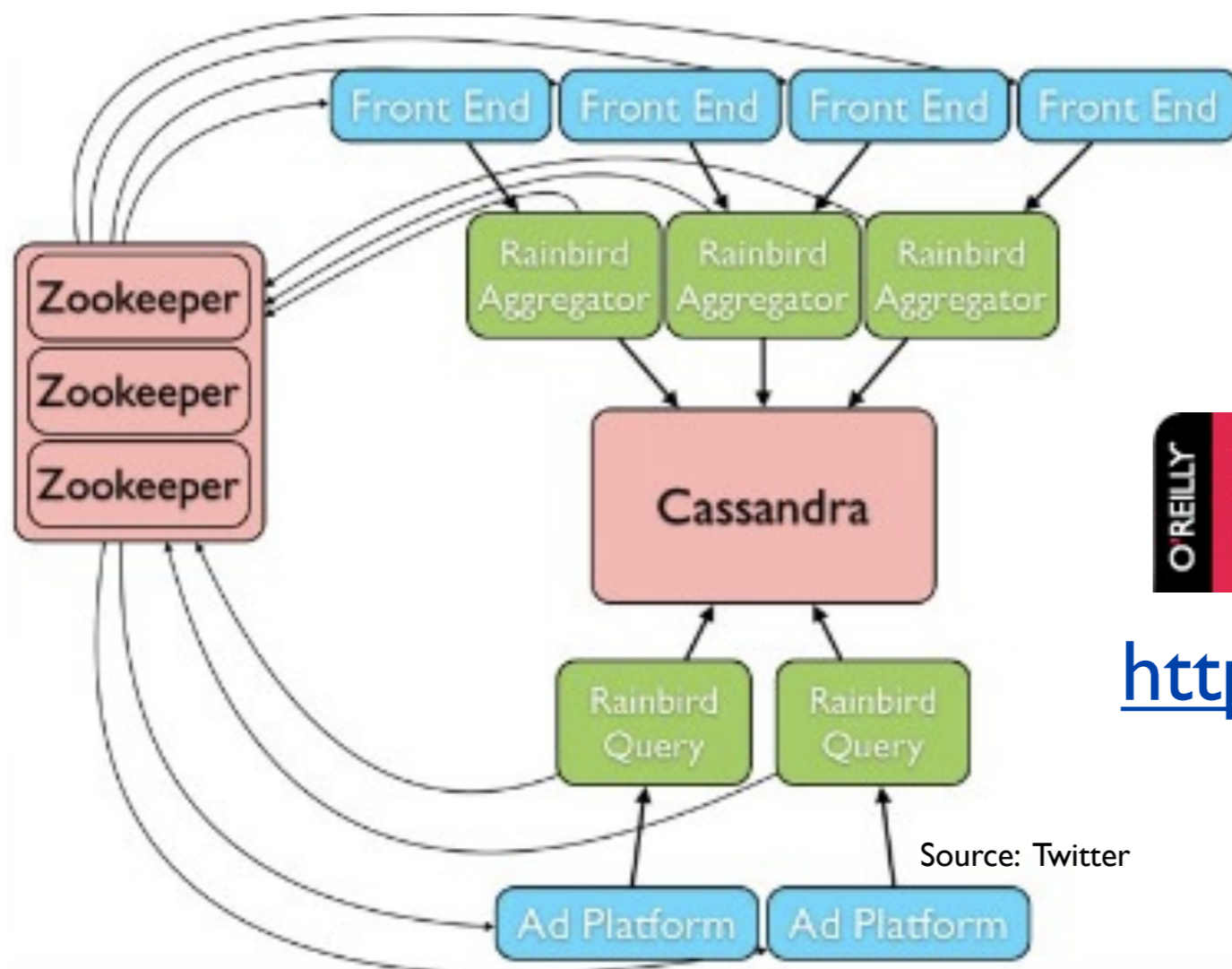


Or... simplify analytics



Source: Twitter

Rainbird: analytics on +1s



Twitter



<http://goo.gl/C945w>

‘Collect’ and ‘Serve’ stages like a single DB client

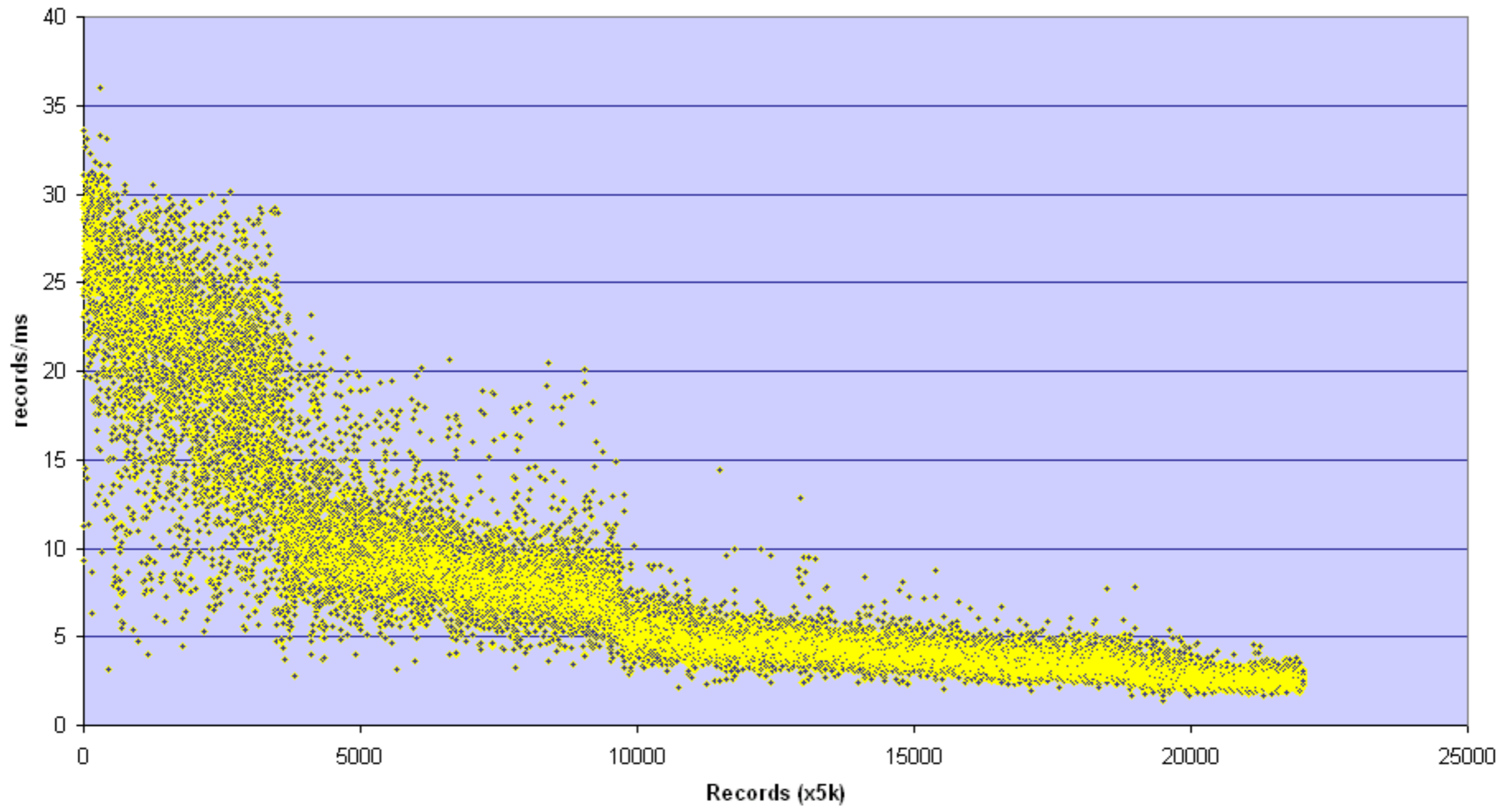




Let's unify the pipeline

- Ingest huge semi-structured workloads
- Run compute inside the database
- Double as a high-availability serving layer
- One platform, many interfaces

Why not RDBMS?



Source: Gerard Maas, <http://www.gerardmaas.net/2011/06/bigdata-on-rdbms>

Why not Hadoop?

- Map Reduce: batch-oriented, > 10s cycle time
- Fixed inputs: forces arbitrary boundaries
- No processing until batch fully loaded
- Hadoop Online Project: research phase
- Drawn To Scale: indices to speed retrieval

Rich, scalable databases



- High write throughput
- ‘Coprocessors’ basis for incremental analytics?
- Great Hadoop integration
- Not high availability: need separate serving tier



Cassandra

- High write throughput
- Atomic counters
- Integrates: Hadoop, Pig, Hive
- Great for serving: high availability, cross DC

Acunu

- Cassandra distro with management tools
- Castle: in-kernel native key-value store
- 'BerkeleyDB and ZFS for Big Data Pipeline'



100x Cassandra
performance

4x Cassandra
+ SSD optimization

Ask yourself...

- Does my problem fit a Pipeline model?
- How valuable are results in X time?
- Dataset size? Raw, cleaned? Rate of change?
- Are approximate results ok?
- How much can the analytics be simplified?
- Where can I get rid of batch work?
- Serving constraints: uptime, latency, #users?

Thanks

Tim Moreton // @timmoreton



www.acunu.com @acunu

<http://www.flickr.com/photos/cmpalmer/99806770/>

<http://www.flickr.com/photos/o5com/5488964999/>

<http://www.flickr.com/photos/wyldkyss/3480600874/>

Apache, Apache Cassandra, Cassandra, Hadoop, HBase, and the eye and elephant logos are [trademarks of the Apache Software Foundation](#).