

Riak to the Rescue

Migrating Big Data



Big Data.

Buzzwords.

Don't
believe the
Hype.



Who am I?

Support
Development
SysAdmin
Managing Operations



Operations

8 Ops Engineers
4 Offices



Operations

650 physical
200 virtual
3 data centres

Contact

- Based in Berlin
- twitter: @geidies
- seb@meltwater.com
- <http://underthehood.meltwater.com/>

Migrating Big Data

- Meltwater
- Social Media Data Volumes
- Try and Fail
- Analyse and Succeed
- Things to Learn

Meltwater

INTEGRATED

Meltwater

News Monitoring

Paper-Clip

Read News
Cut and Glue
Telefax

Meltwater News

Crawl the Web
Match new Articles
Morning Report
Analytics UI

Products

PR

m|news
m|press

Marketing

m|buzz / engage
icerocket

SaaS

Subscription model
24,000 clients



basho

basho

riak

- Open Source
- Dynamo Paper
- Erlang



2.0

OMG, OMG!!

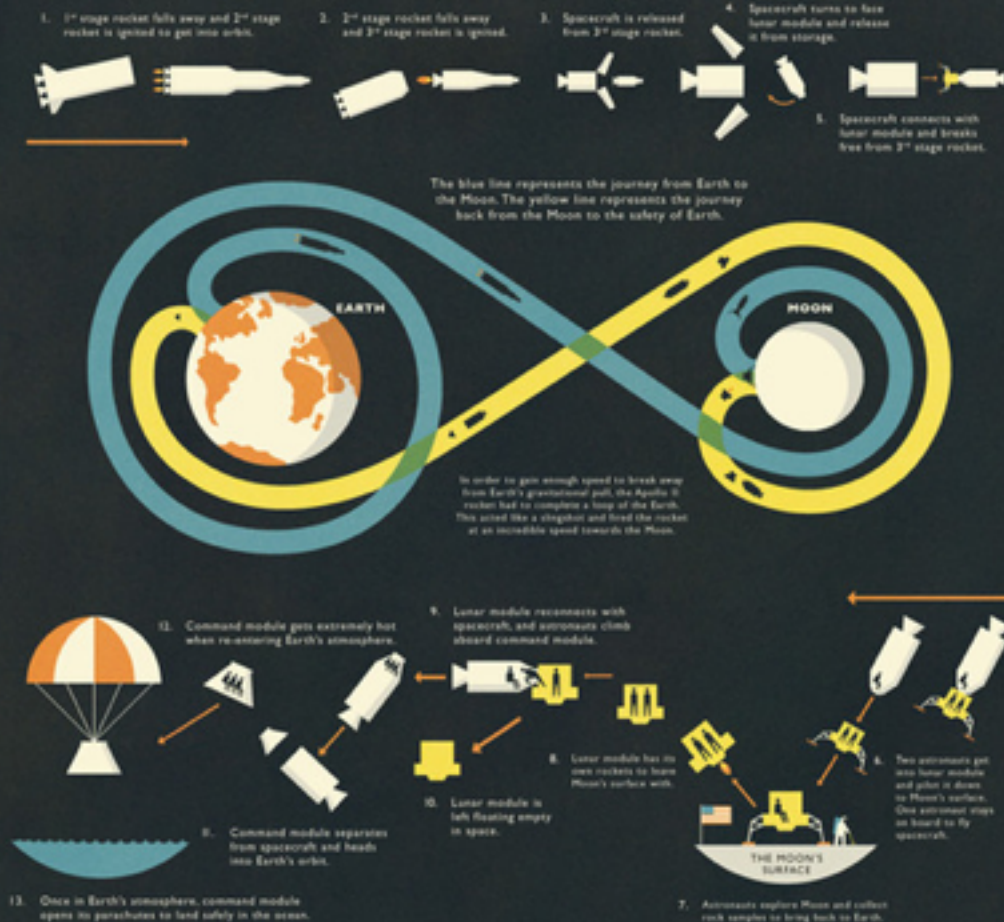
thanks, basho.

Meltwater Buzz

INTEGRATED MARKETING

GOING TO THE MOON

The first successful landing on the Moon took place in 1969, and the journey took the Apollo 11 astronauts 3 days. The following diagram illustrates how they got there and back home safely.



m|news
m|buzz

20 D/s - 8400 S/s
600 D/s - ??

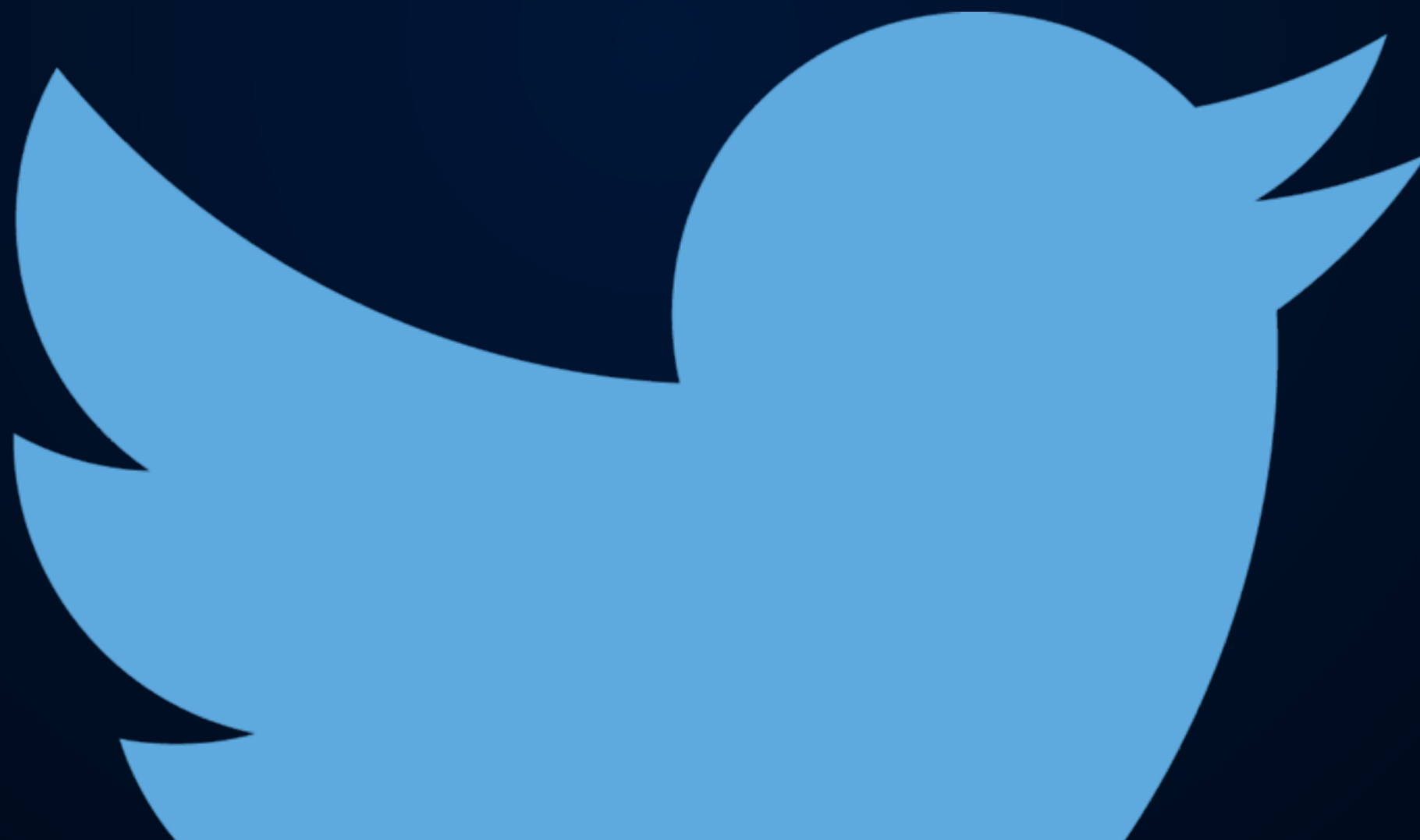
Interesting Shtuff

By Joan Doe - 2014/05/06

Something amazing happened yesterday. It was more interesting than what happened the day before, but maybe it won't change the events that are about to come tomorrow.

What does Lorem ipsum dolor really mean? we know it is not real latin. But it looks pretty good, since the characters are evenly distributed. I once tried translating it, and it really doesn't make any sense. Talking here is amazing. Wow, Denmark - it's actually really cool being in Aarhus. You should have a chat with me after the talk if you have further questions. Please don't hesitate to say hi. If you're in Berlin, come stop by the meltwater office for a chat about big data, a cup of coffee, a game of table tennis or foosball. You can find us at Rotherstraße 22 in Friedrichshain.

You can find us at Rotherstraße 22 in Friedrichshain. data, a cup of coffee, a game of table tennis or foosball. come stop by the meltwater office for a chat about big. Please don't hesitate to say hi. If you're in Berlin,



Social Media

- 140 Characters
- Pages Long

Social Media

- Metadata
 - Location
 - Followers
 - Threads

Social Media

- Extracted Metadata
 - sentiment
 - named entities
 - intent
- Editorial vs. Opinion vs. Both

m|buzz version 1

- Buzzgain
- php, MySQL, SolR



Attention!

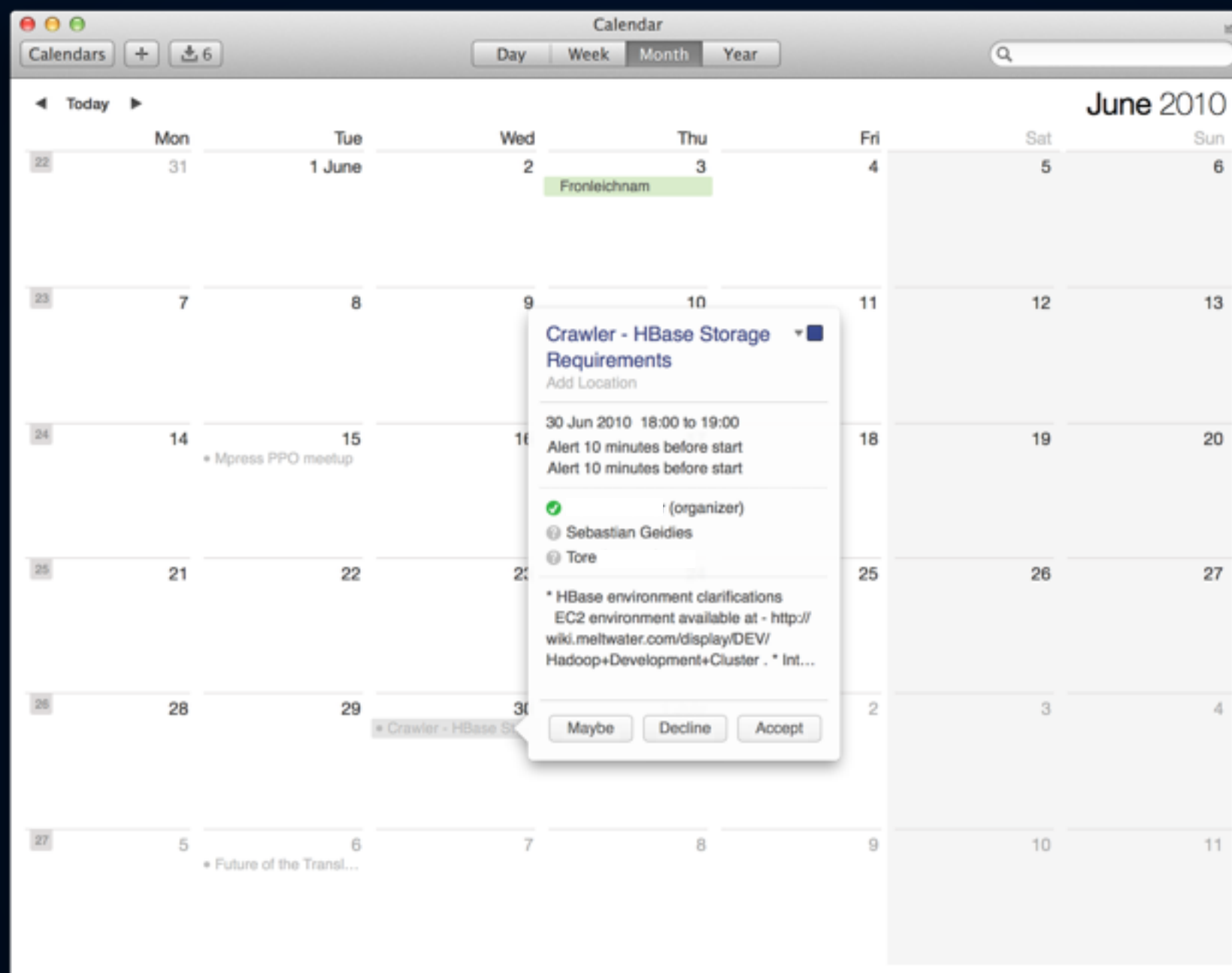
Attention!

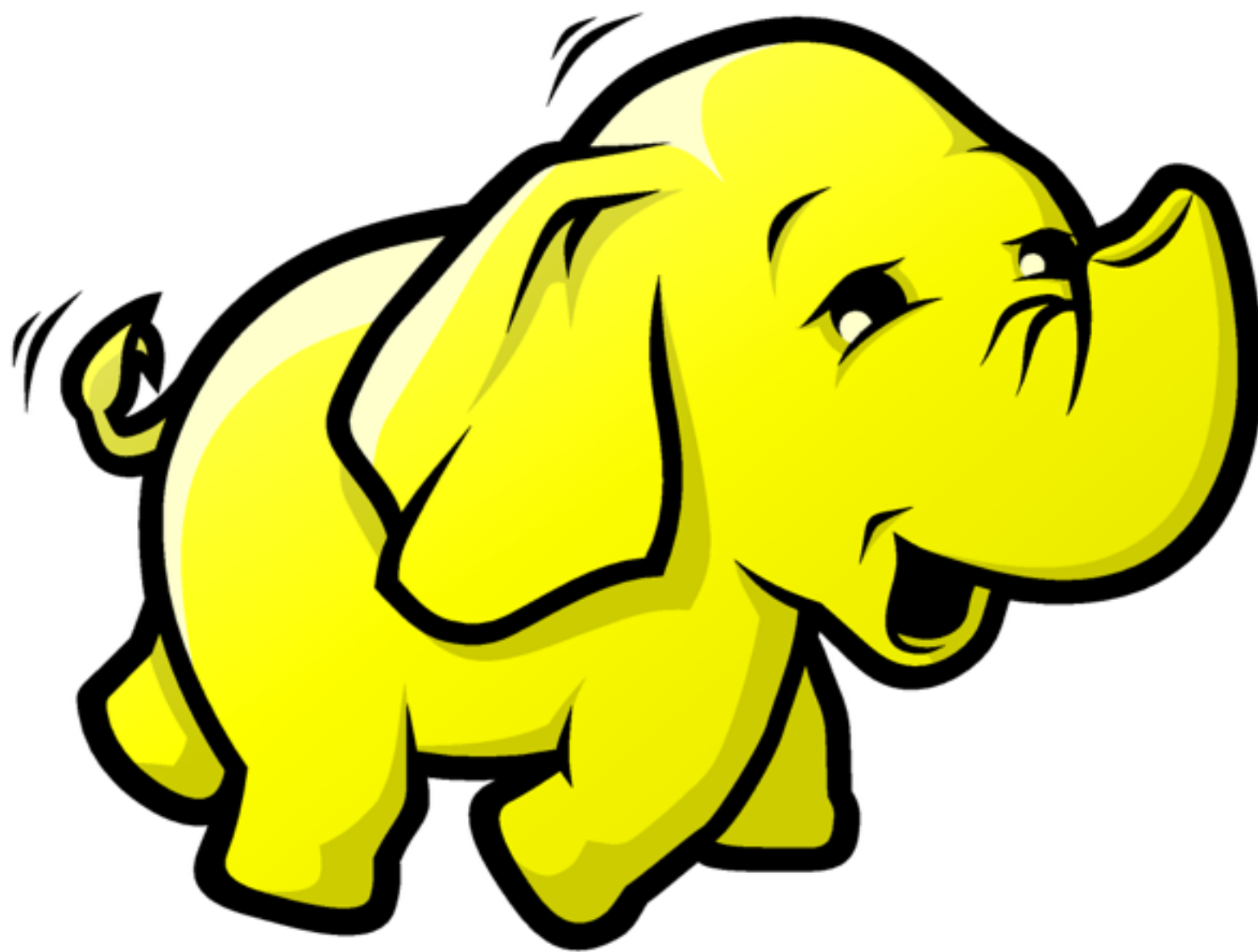
Your Use Case

Research
Evaluate
Test

m|buzz version 2

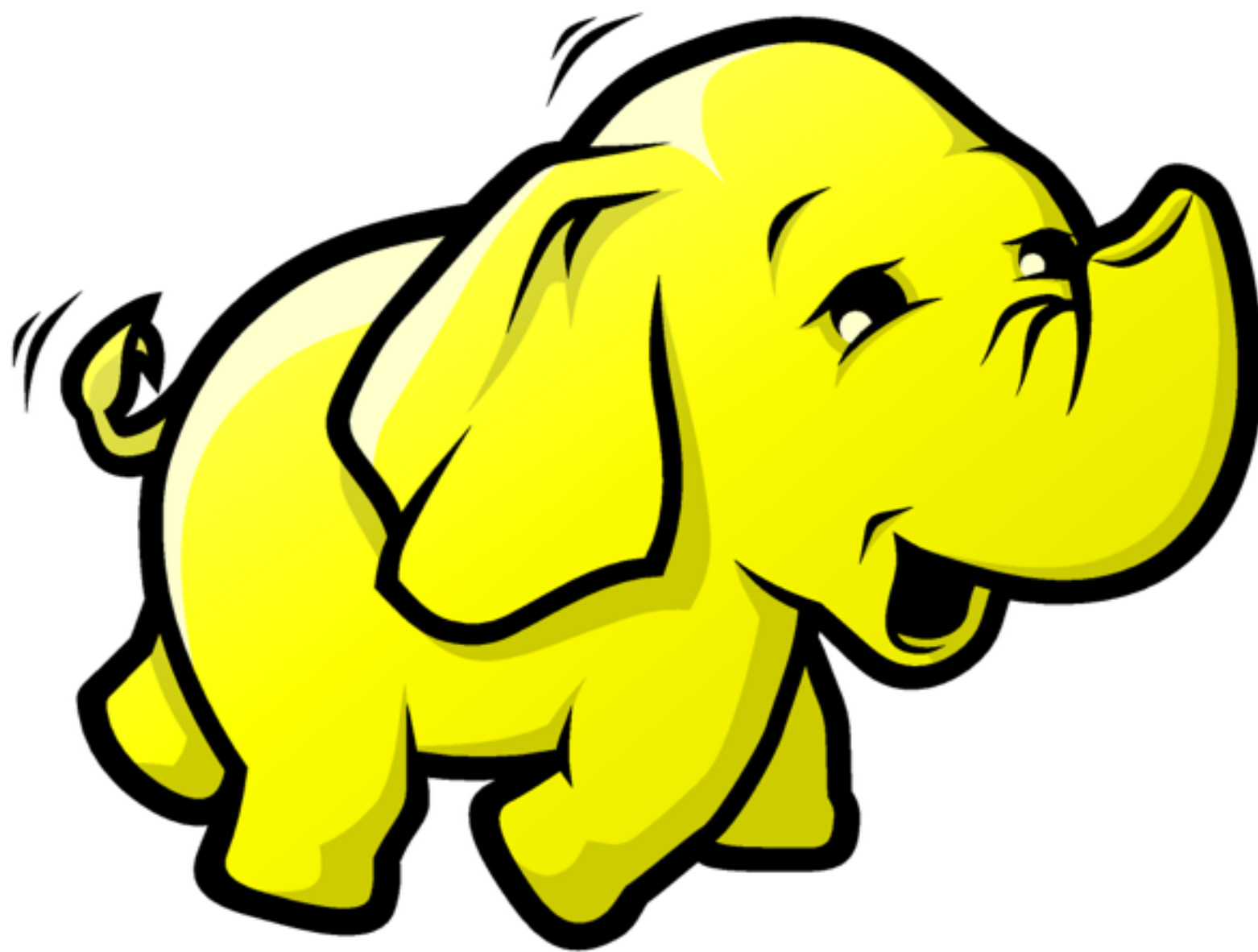
Scalability,
Features,
Buzzwords!





“Some people, when confronted with a problem,
think “I know, I’ll use regular expressions.” Now
they have two problems.”

– *Jamie Zawinski*



Requirements

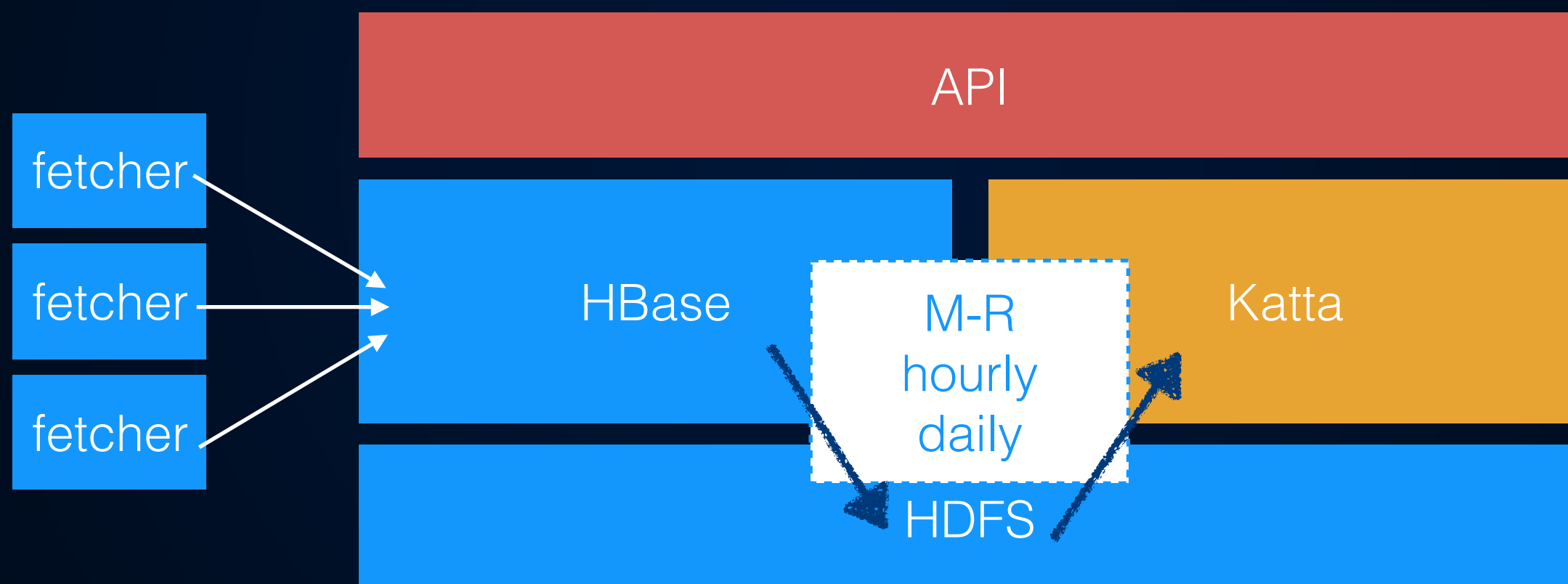
- Fail-Safety
- High Availability
- A Lot of Unstructured Data
- Near-Real-Time Indexing
- Time-Based Ordering instead of Relevancy

m|buzz version 2

- Hadoop Ecosystem
- Apache Projects

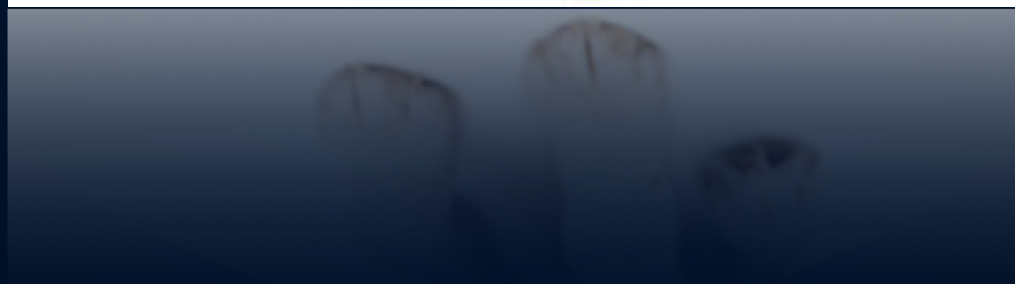
The logo for Apache HBase. The word "APACHE" is in a light gray, sans-serif font. Below it, the word "HBASE" is in a bold, red, sans-serif font.

m|buzz version 2



It's a trap!

- buzzwords
- commodity hardware
- scale







- Build upon lucene
- Master -> Worker -> Client
- communication through zookeeper
- multiple index copies
- copied from HDFS -> local disk





- OK in theory.
- Out Of Memory
- Garbage Collection Hell
- version 0.62 - odd bugs.





on May 26, 2011  **0.6.4** ...
- 9630067

on Nov 24, 2010  **0.6.3** ...
- 96dcd07

on Aug 2, 2010  **0.6.2** ...
- 46c2b5e

on Feb 28, 2010  **0.6.1** ...
- d95891c

on Feb 4, 2010  **0.6.0** ...
- 9953b47

on Jan 12, 2010  **0.6.rc1** ...
- 606039a

on May 26, 2011

on Nov 24, 2010

on Aug 2, 2010

on Feb 28, 2010

on Feb 4, 2010

on Jan 12, 2010

sgroschupf /

Releases Tags

programming coding lambda

geidies + -

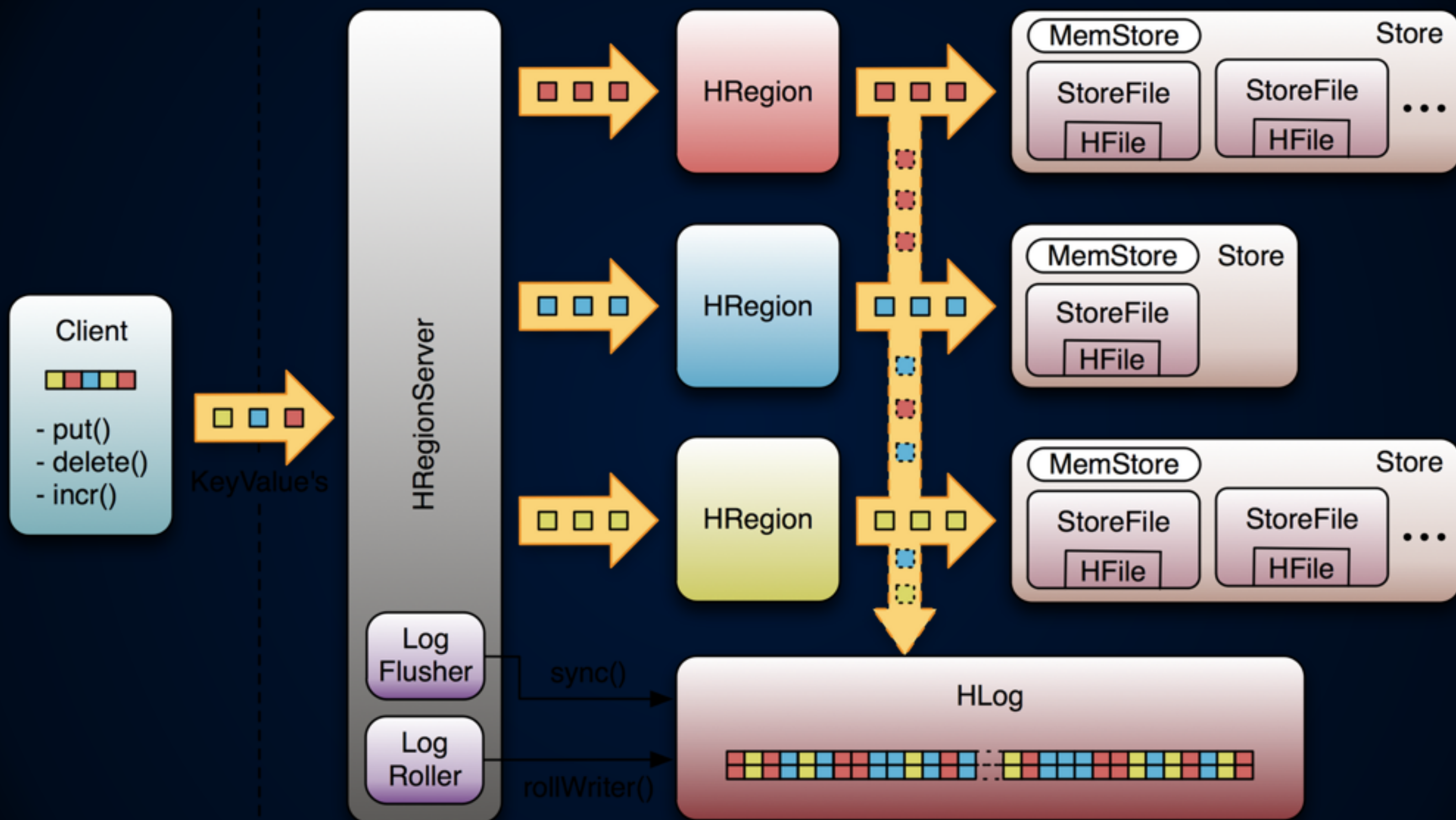
Star 27 Fork 9

Reader

APACHE
HBASE

0205

APACHE
HBASE



keyspace

split



key a -> key c

key n -> key o

-ROOT-

.META.

Fail-Safety

Fail-Safety

Does NOT mean High Availability

Data on a Single Node

Minutes.

55,000 posts / minute

Funny Regions

Overlapping
Gaps
Negative Length

Funny Regions

```
REGION => {NAME => 'buzz_data',  
1333073443000_62gfsHBsE5vNSz168ByvP5tDPu0A,1333173530871',  
STARTKEY => '1333073443000_62gfsHBsE5vNSz168ByvP5tDPu0A',  
ENDKEY => '1326306499000_evKK670FSV9MAas2CMZAr41wLm0A', ENCODED =>  
128988498, TABLE => {{NAME => 'buzz_data', FAMILIES => [{NAME =>  
'fm_contents',VERSIONS => '1', COMPRESSION => 'LZO', TTL => '2147483647',  
BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}, {NAME =>  
'fm_input_info', VERSIONS=> '1', COMPRESSION => 'LZO', TTL => '2147483647',  
BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}, {NAME =>  
'fm_metadata', VERSIONS => '1', COMPRESSION => 'LZO', TTL => '2147483647',  
BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}, {NAME =>  
'fm_output_info', VERSIONS => '1', COMPRESSION => 'LZO', TTL => '2147483647',  
BLOCKSIZE => '65536', IN_MEMORY => 'false', BLOCKCACHE => 'true'}}]}
```

HBase

- .META. corruption
- Data Unavailability
- Slow Start of Regions
- Full Cluster Restarts Slow
- Hotspots

Good News!

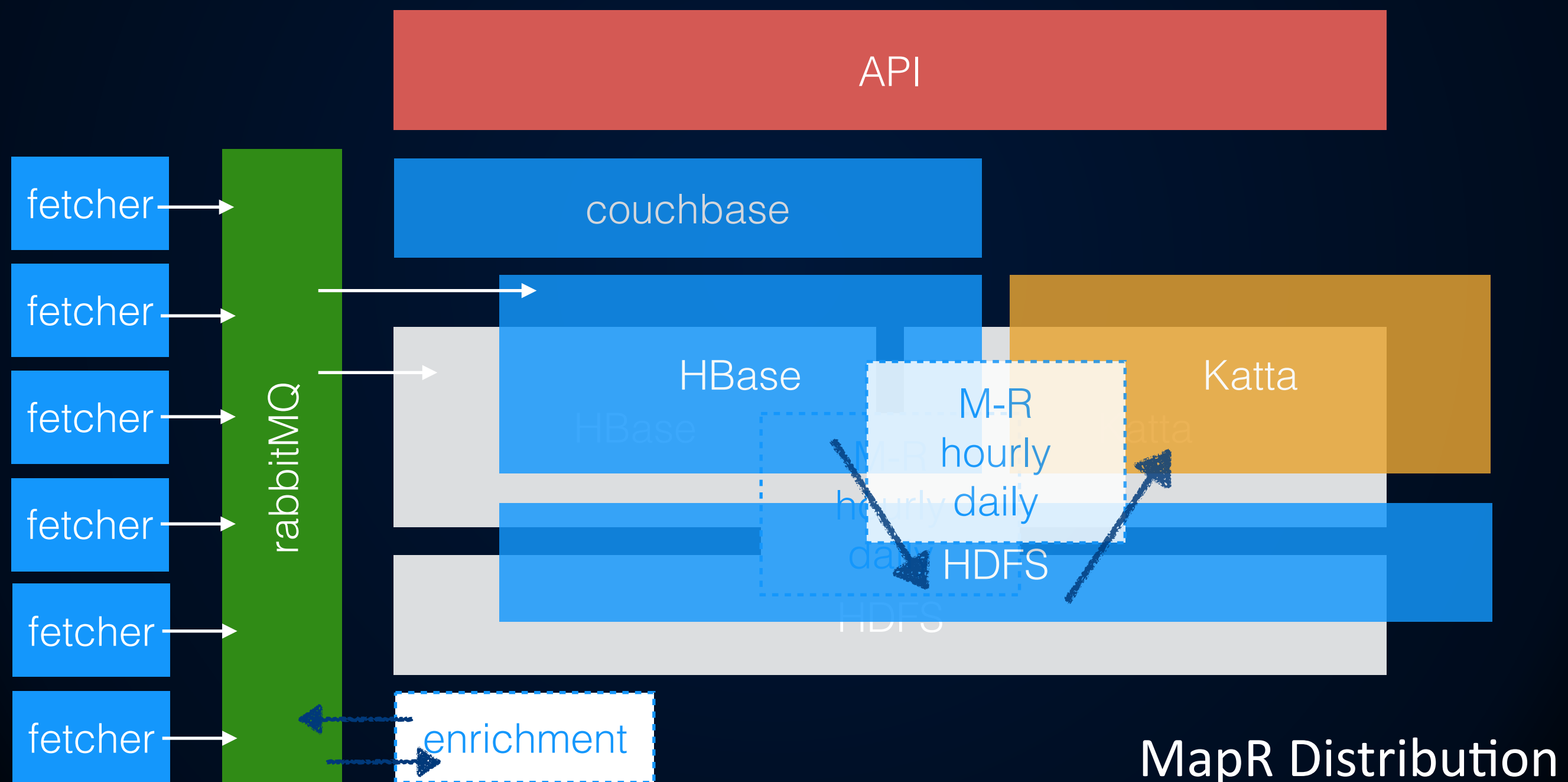
NameNode never crashed.
Great.

Changes...

...do you speak it?

MAPR®

m|buzz version 2.5



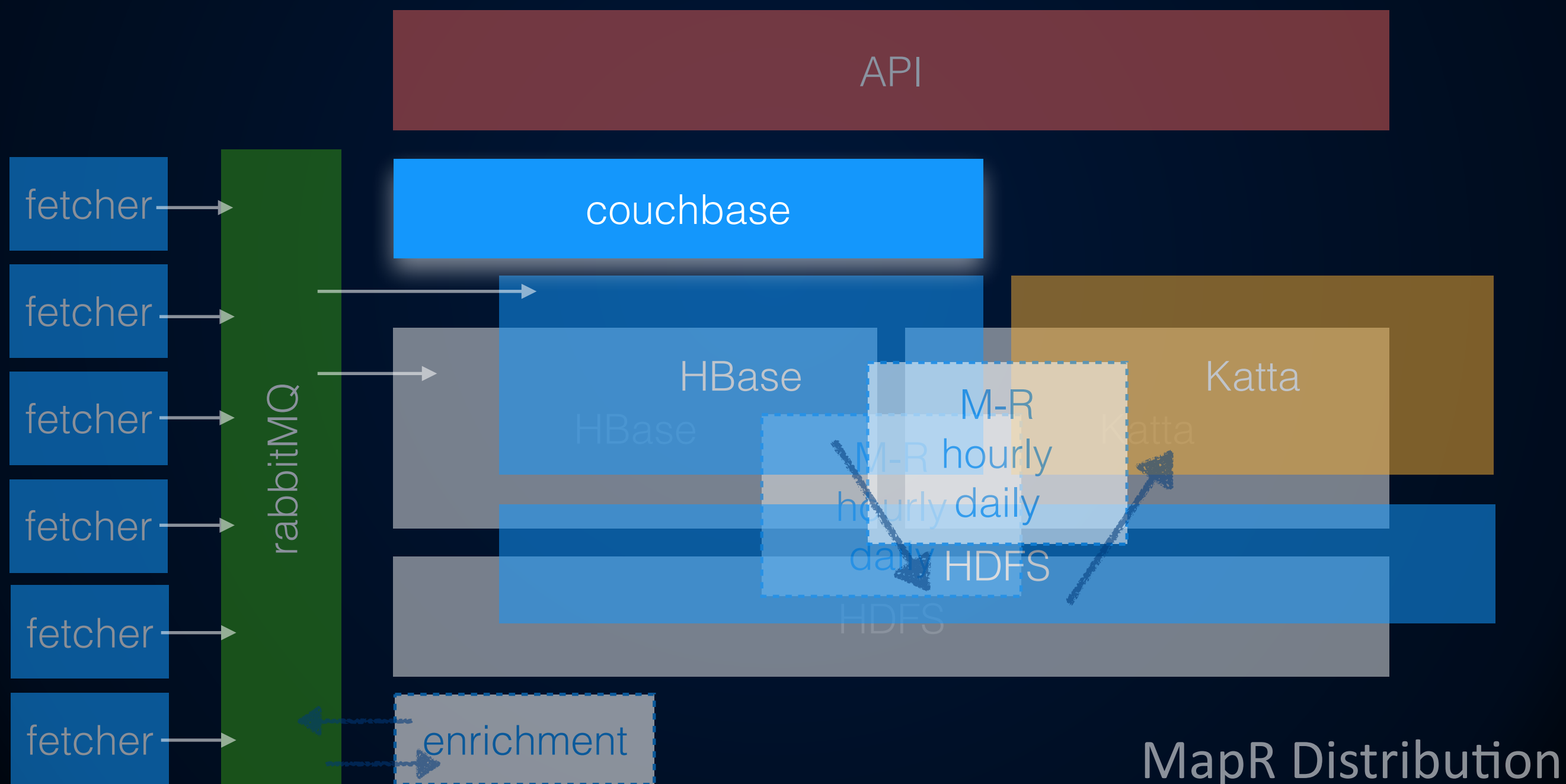






- Message Queue System
- Erlang
- Redundant Setup, fail-safe and high-available
- Write to Exchange -> Distribute to Multiple Queues

m|buzz version 2.5



Couchbase



First Read Wins

Parallel Reads:

couchbase

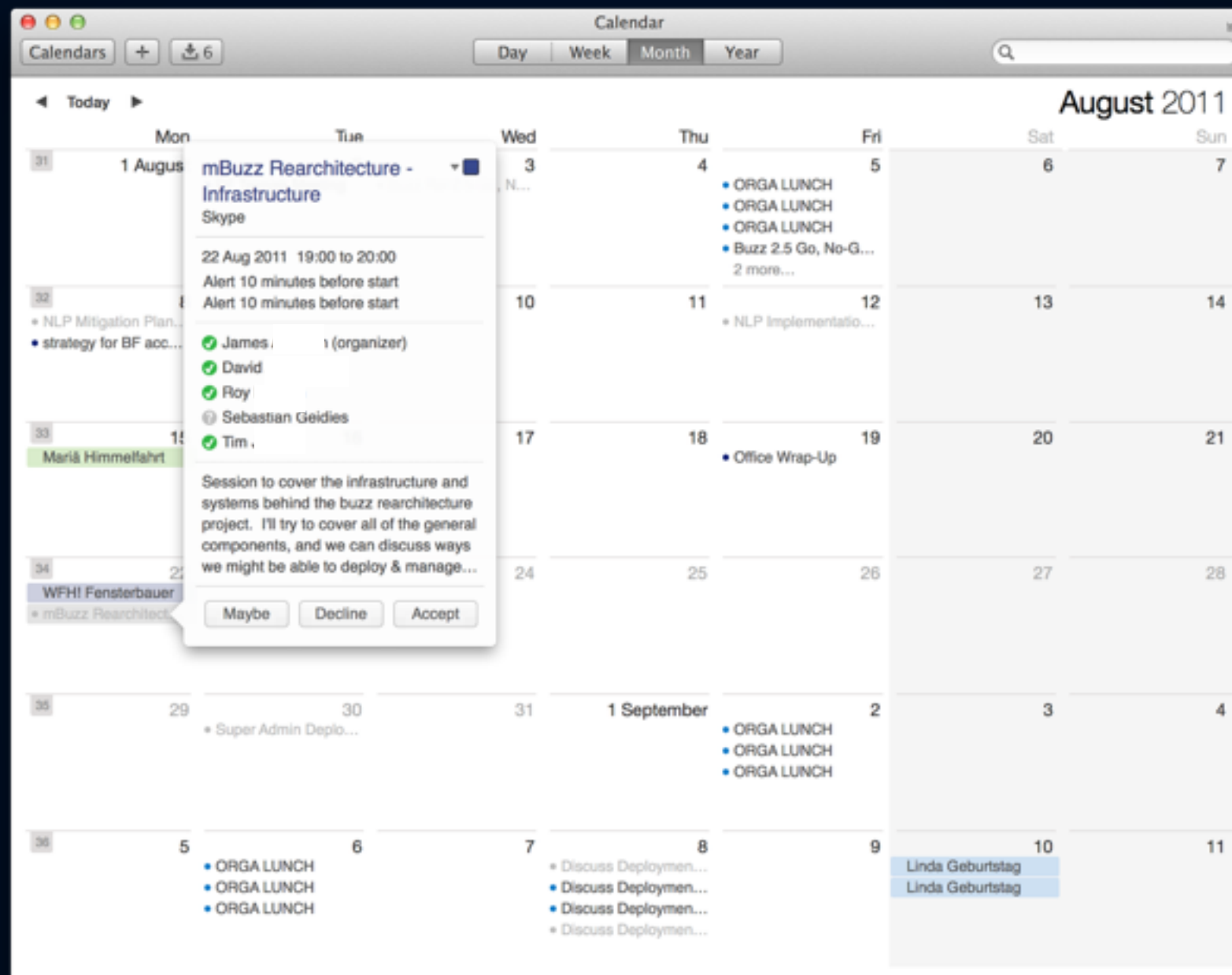
vanilla HBase

MapR HBase

couchbase scales!

...to four weeks of data.
2.2B entries
TTL

Are we there yet?



• ORGA LUNCH
• ORGA LUNCH
• ORGA LUNCH

• Discuss Deploymen...
• Discuss Deploymen...
• Discuss Deploymen...

Linda Geburtstag
Linda Geburtstag

Options

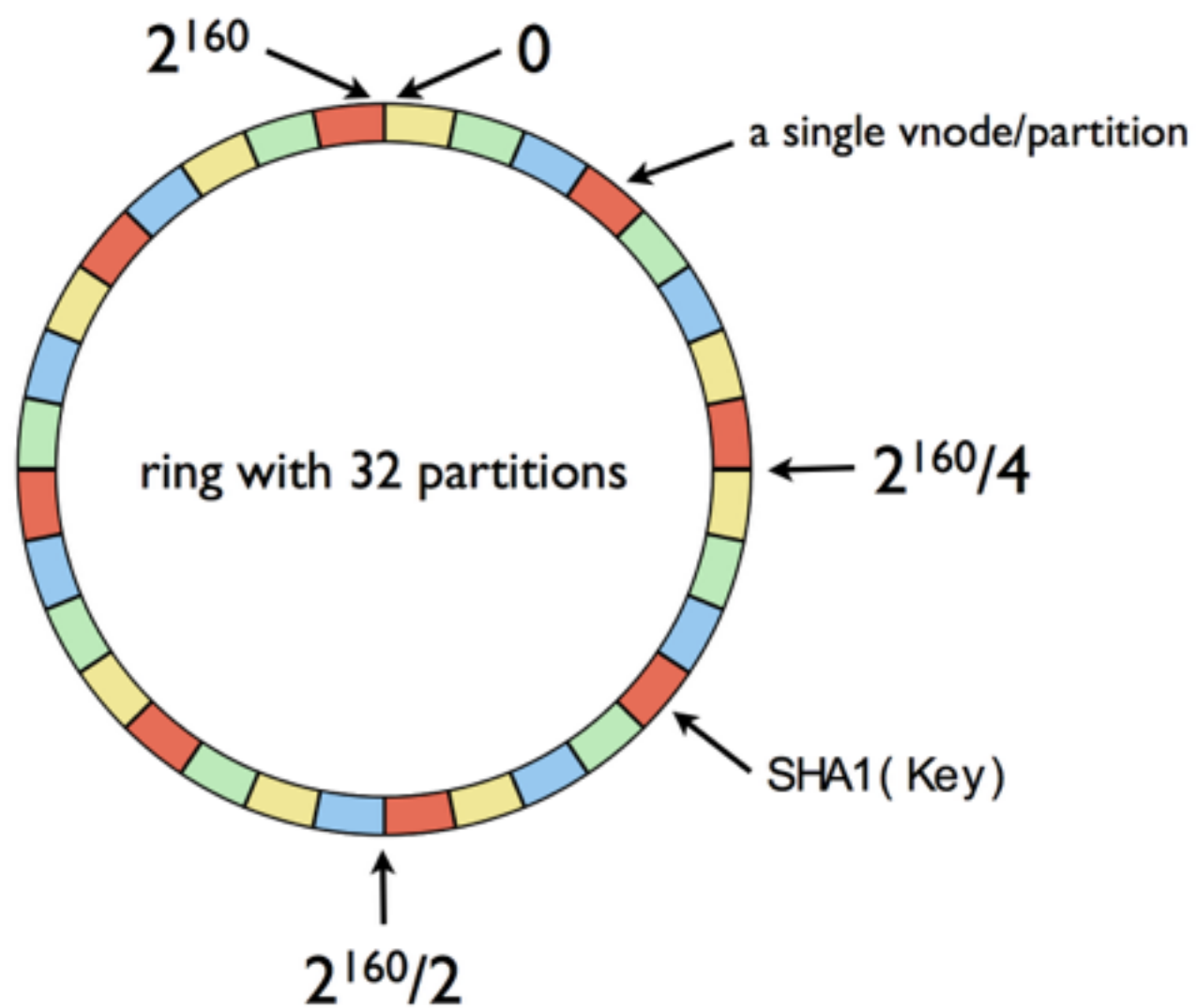
	Pro	Con
custom WAL	works safely	doesn't scale (easily)
MySQL cluster	A lot of experience	hitting limit of scaling
commercial Object storage	commercial support	up-front investment
riak		

Requirements

- ✓ High Availability
- ✓ Data Safety
- ✓ Scalability
- ? Range Scans or TTL to limit data

riak

Key-Value model
Objects in
Buckets



Node 0

Node 1

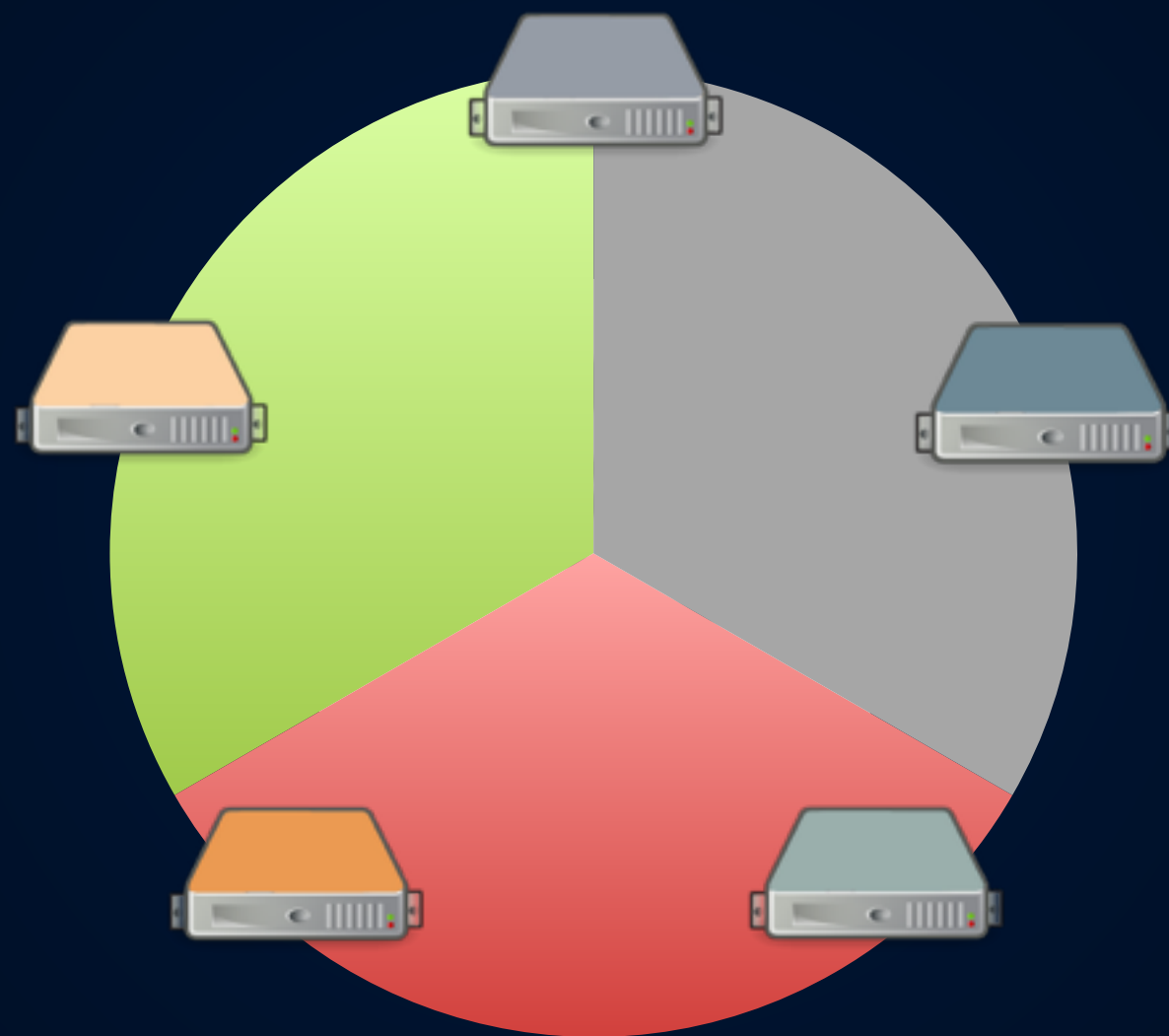
Node 2

Node 3

$\Sigma_{1 \in \mathcal{O}} \mathcal{V}$



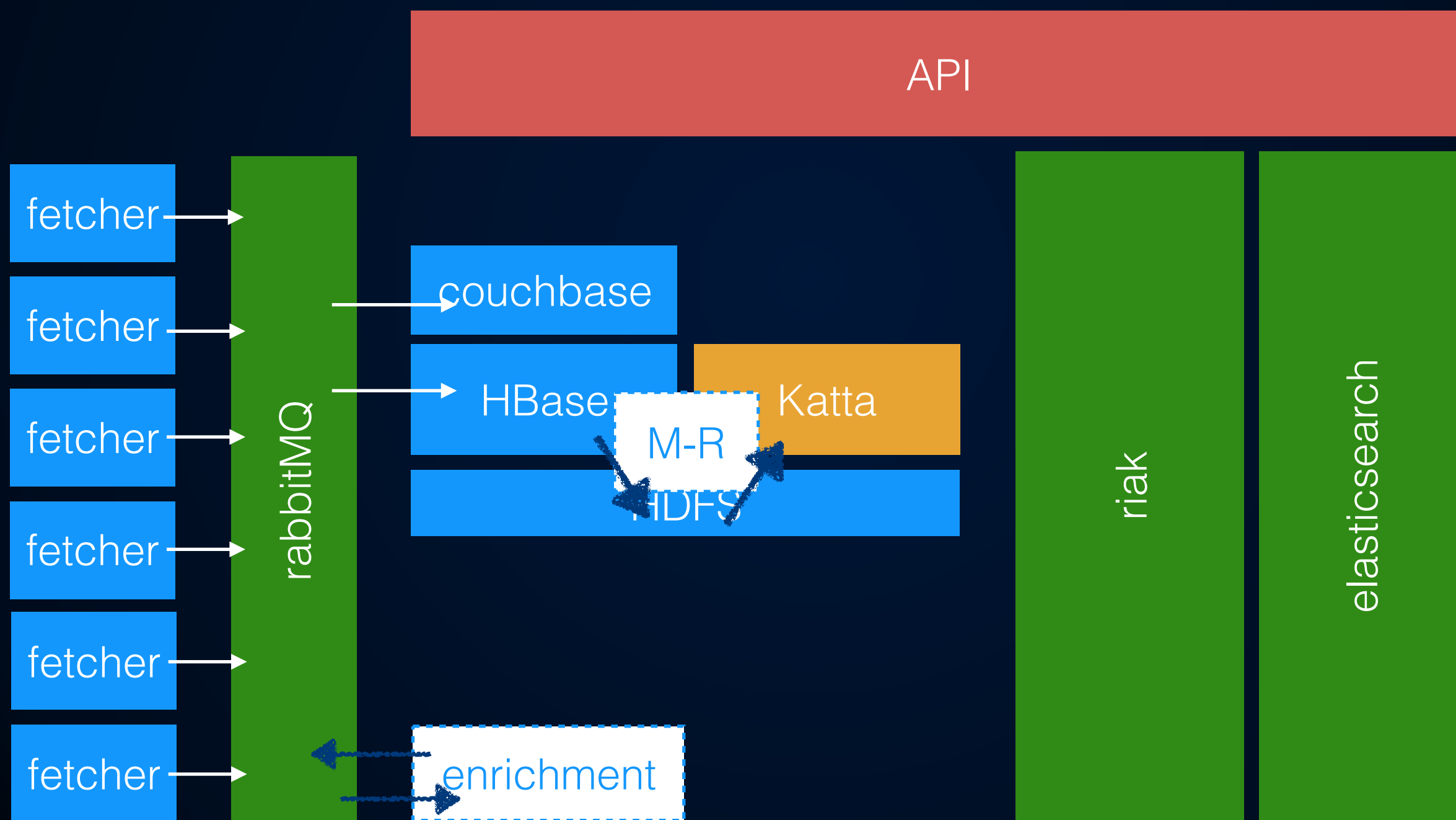




“While there are mechanisms such as Vector Clocks to help deal with these issues, if your application requires the kind of strong consistency found in ACID systems, Riak may not be a good fit.”

– *riak documentation*

m|buzz version 2.6



Commodity Hardware

- HP DL360 G1
- 4c CPU
- 32GB RAM
- 1x 2TB 7.2k spinner
- ...37 of those.

Configuration

- levelDB
- erlang VM
- Map-Reduce

Future-Proof

Setting the ring-size to...

2048.

“2048 is definitely the upper bound of what we recommend, but with the right amount of machines, this can work.”

– *riak mailing list*

“Are you guys insane? We didn’t even know that
was possible!!”

– *riak mailing list re-niced*

Numbers

- 37 nodes
- 55,000 writes per minute
- 350,000 reads per minute
- 1.8TB data per node

Hey, wait.

A good three weeks?

Let's do it.

parallel reads
gather numbers
stability
speed

riak is slow.

but consistent,
and massively parallel.

~~riak is slow.~~

riak is not as fast as a memory-only
key-value store.

stability over
speed.

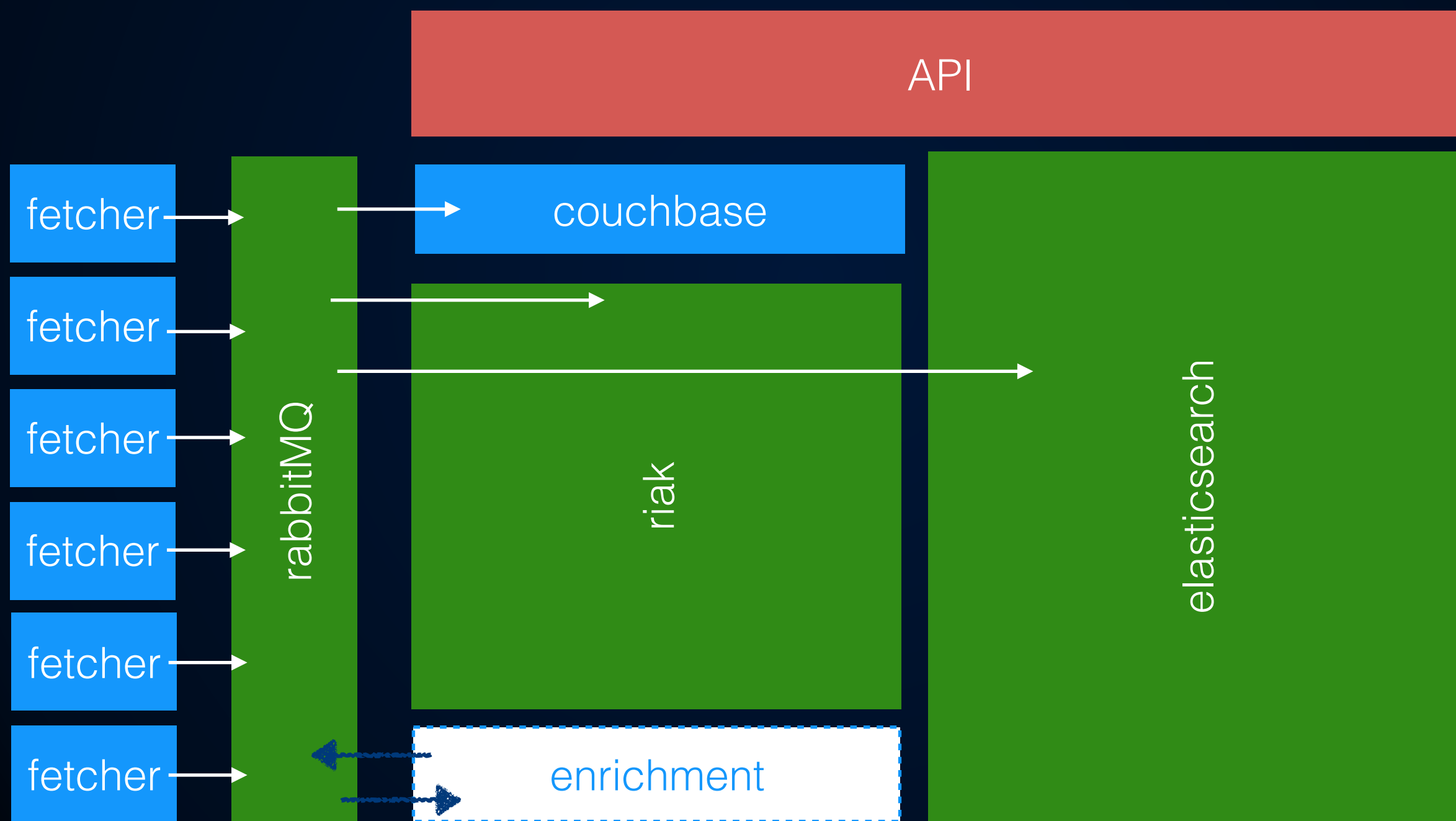
stability

- availability during
 - node failures
 - upgrades
 - configuration updates

Search

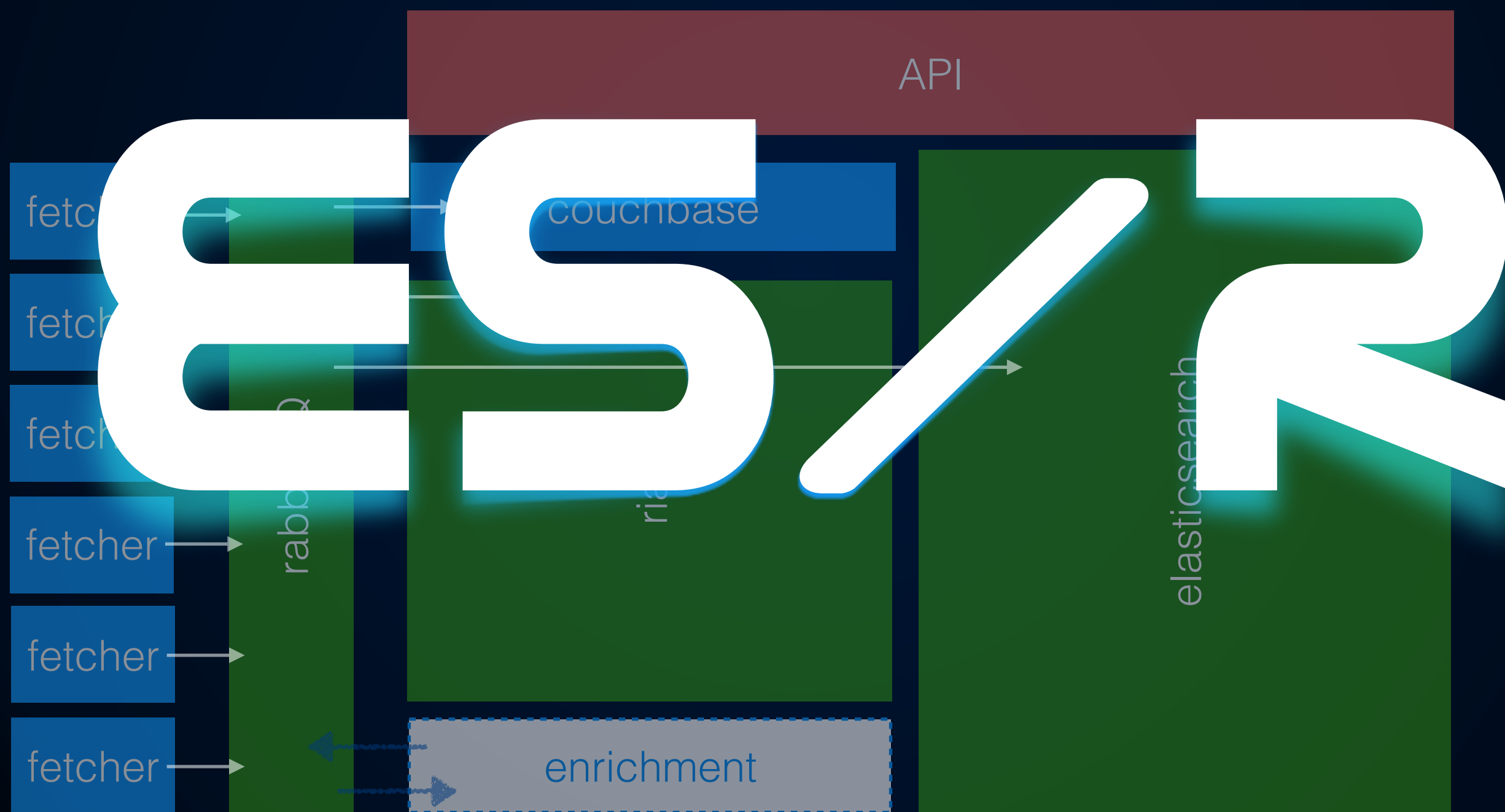
Search

m|buzz version 3



Naming Things

m|buzz ~~version 3~~



Putting it live

10:20 AM

Adarsh T. checked out few accounts looks good



Stian G. performing very well indeed



Adarsh T. true :-)



10:25 AM

Sebastian G. Ok. So here's the truth: for the past hour, we've been running of ES/R instead of the 2.2 API.



There is some side effects. Example given: I've got posts in my campaign that are only 2 minutes old.



:)



Matthias R. Tststs... sneaky bastard ;-)



Sebastian G. I wanted people to test unbiased. :)



Still live

- 58,000,000,000 key-value pairs written
- 365,000,000,000 reads
- 3.5ms mean (8ms 95th, 35ms 99th, 2s 100)

Monitoring

- Input “valves”
- throughput of any intermediate processing step
- output valves
- distribution of data across cluster
- handovers of data within the cluster

Dashboards

And APIs.

The screenshot shows a web browser window titled "localhost Hadoop Map/Reduce Administration". The address bar shows "http://localhost:50030/jobtracker.jsp". The page has a "Hadoop Tracker" tab and a "Quick Links" link in the top right. The main heading is "localhost Hadoop Map/Reduce Administration". Below this, the status is "State: INITIALIZING", with details: "Started: Fri Jun 25 14:34:57 PDT 2010", "Version: 0.20.2-dev, r", "Compiled: Fri Oct 23 21:54:18 PDT 2009 by jherr", and "Identifier: 201006251434". A section titled "Cluster Summary (Heap Size is 81.06 MB/995.88 MB)" contains a table with the following data:

Maps	Reduces	Total Submissions	Nodes	Map Task Capacity	Reduce Task Capacity	Avg. Tasks/Node	Blacklisted Nodes
0	0	0	0	0	0	-	0

Below the table is a "Scheduling Information" section with a table:

Queue Name	Scheduling Information
default	N/A

There is a "Filter (JobId, Priority, User, Name)" input field with an example: "Example: 'user which 2222' will filter by 'which' only in the user field and '2222' in all fields". The "Running Jobs" section shows a "none" button. The "Completed Jobs" section is empty.

Master: ip-10-110-247-97.ec2.internal:60000

[Local logs](#), [Thread Dump](#), [Log Level](#), [Debug dump](#)

Attributes

Attribute Name	Value	Description
HBase Version	0.92.0, r1234893	HBase version and revision
HBase Compiled	Sun Feb 19 02:45:09 PST 2012, gkesavan	When HBase version was compiled
Hadoop Version	1.0.1, r1243785	Hadoop version and revision
Hadoop Compiled	Tue Feb 14 08:12:14 UTC 2012, hortonfo	When Hadoop version was compiled
HBase Root Directory	hdfs://ip-10-140-6-87.ec2.internal:8020/apps/hbase/data	Location of HBase home directory
HBase Cluster ID	b3d30236-4f60-4a39-ac40-76d5ffac6455	Unique identifier generated for each cluster
Load average	◆	Average number of regions per region server
Zookeeper Quorum	domU-12-31-39-0A-06-14.compute-1.internal:2181,ip-10-46-197-123.ec2.internal:2181	Addresses of all registered ZK servers
Coprocessors	[]	Coprocessors currently loaded locally
HMaster Start Time	Tue Feb 21 19:18:24 EST 2012	Date stamp of when this HMaster started
HMaster Active Time	Tue Feb 21 19:18:24 EST 2012	Date stamp of when this HMaster became active

Tasks

[Show All Monitored Tasks](#) [Show non-RPC Tasks](#) [Show All RPC Handler Tasks](#) [Show Active RPC Calls](#) [Show Client Operations](#) [View as JSON](#)

Start Time	Description
Tue Feb 21 19:27:10 EST 2012	Doing distributed log split in [hdfs://ip-10-140-6-87.ec2.internal:8020/apps/hbase/data/.logs/ip-10-190-187-12.ec2.internal,60020,1329103193864-splitting]

Tables

Catalog Table	Description
-ROOT-	The -ROOT- table holds references to all .META. regions.
.META.	The .META. table holds references to all User Table regions


1 table(s) in set. [\[Details\]](#)

User Table	Description
usertable	{NAME => 'usertable', FAMILIES => [{NAME => 'family', MIN_VERSIONS => '0'}]}

RabbitMQ Management

[rabbit-ops.melt.no/#/queues/%2F/pipeline](#)

[Pin It](#)
[SoundCloud Download](#)
[MNews stack](#)
[Passpack It!](#)
[DuckDuckGo](#)
[Wikipedia](#)
[News](#)
[Beliebt](#)
[Oto255](#)
[programming](#)
[coding](#)
[lambda](#)



User: Administrator
RabbitMQ 3.1.5, Erlang R14B04

Log out


Overview
Connections
Channels
Exchanges
Queues
Admin

Virtual host: /

Queue pipeline

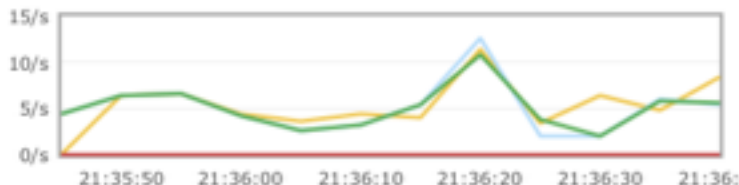
Overview

Queued messages (chart: last minute) (?)



Category	Value
Ready	0 msg
Unacknowledged	8 msg
Total	8 msg

Message rates (chart: last minute) (?)



Category	Value
Publish	4.8/s
Deliver	5.4/s
Redelivered	0.00/s
Acknowledge	5.6/s

Details

Parameters	Status
x-message-ttl: 600000	Active
x-dead-letter-routing-key: pipeline	Consumers 48
x-dead-letter-exchange: com.meltwater.enrichment.deadletter	Memory 1.1MB
durable: true	Virtual host /
Policy ha	Node rabbit@mag-fh-queue02
Exclusive owner None	Mirrors rabbit@mag-fh-queue01


Message rates breakdown


Consumers


Bindings


From	Routing key	Arguments	
(Default exchange binding)			
com.meltwater.enrichment	#		Unbind










 Snapshot


 Cluster


 Ring

 Objects

 Map/Reduce

 Graphs

 Logs

 Support

Ring View

Prev

1

Next

Filter

All Owners

#	Owner Node	KV	Pipe	Search
0	riak@riak3.ack	Active	Active	Fallback
1	riak@riak1.ack	Active	Active	Fallback
2	riak@riak2.ack	Active	Active	Fallback
3	riak@riak4.ack	Active	Active	Fallback
4	riak@riak3.ack	Active	Active	Fallback
5	riak@riak1.ack	Active	Active	Fallback
6	riak@riak2.ack	Active	Active	Fallback
7	riak@riak4.ack	Active	Active	Fallback
8	riak@riak3.ack	Active	Active	Fallback
9	riak@riak1.ack	Active	Active	Fallback
10	riak@riak2.ack	Active	Active	Fallback

necessary but not
sufficient

dashboard

API

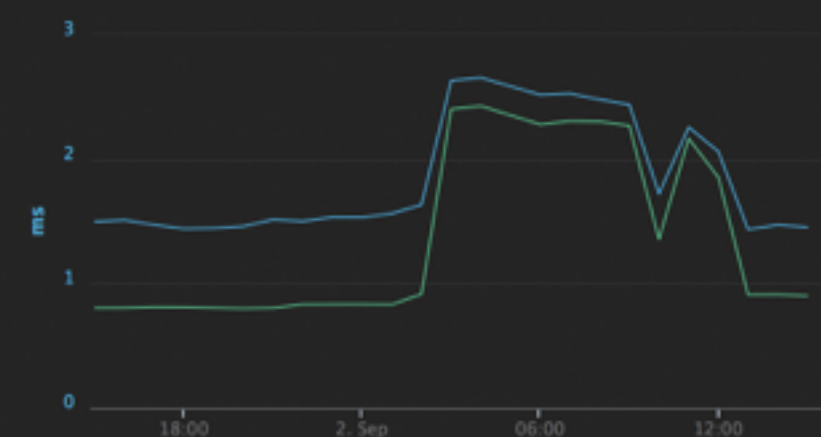
fool-safe performance configuration

good documentation

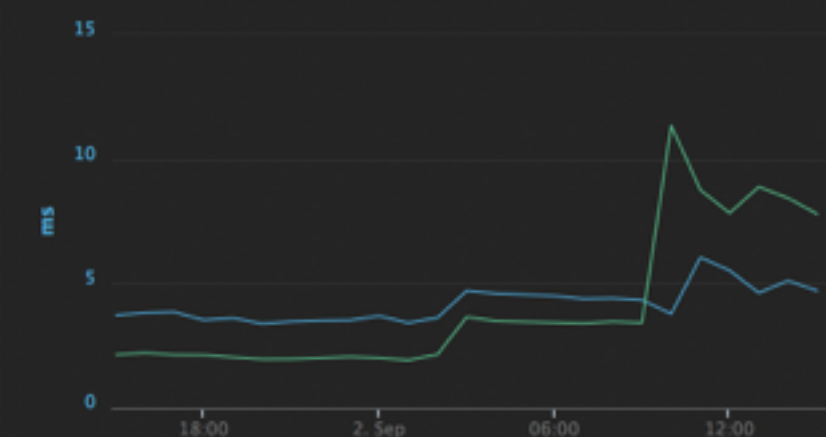
Riak GET/PUT count



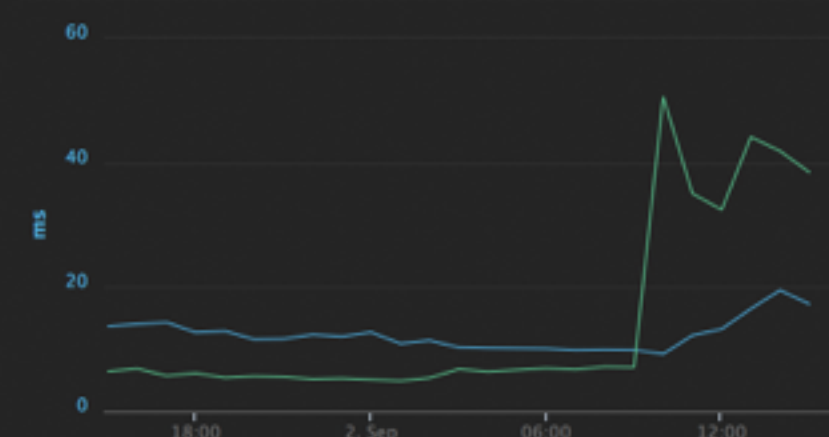
Riak GET/PUT latency (median)



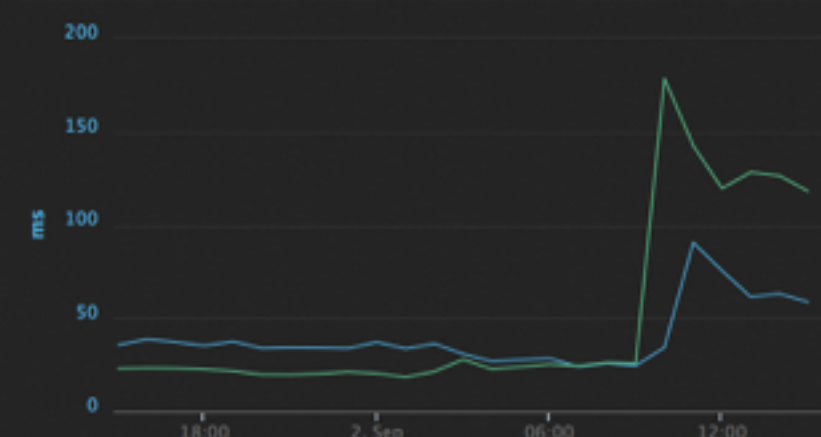
Riak GET/PUT latency (mean)



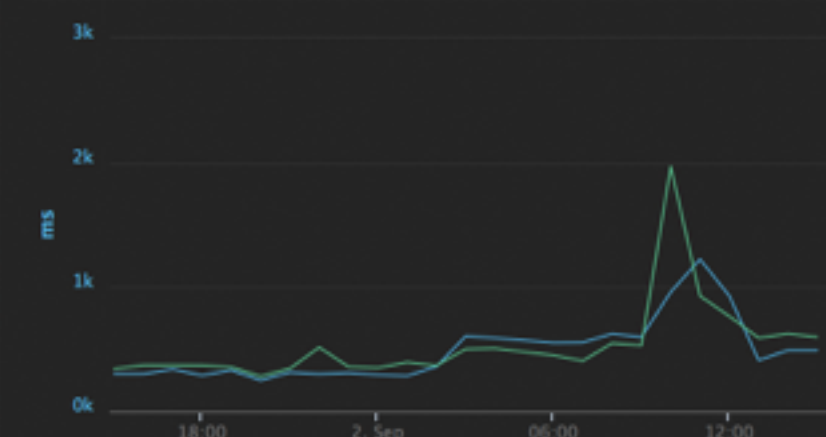
Riak GET/PUT latency (95th)



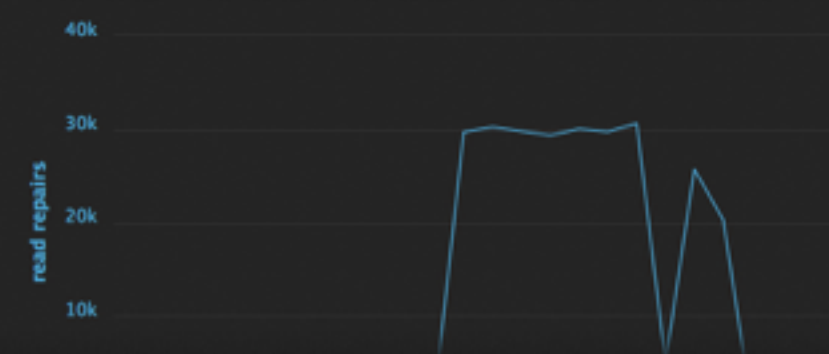
Riak GET/PUT latency (99th)



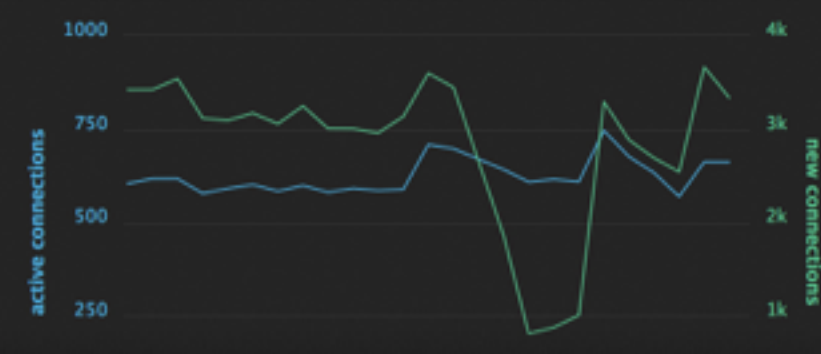
Riak GET/PUT latency (100th)



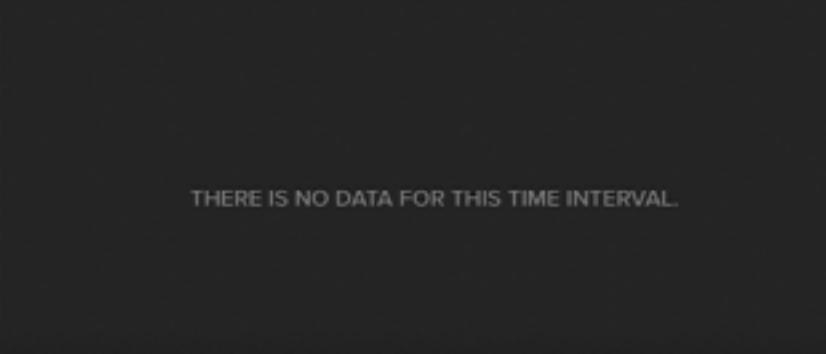
Riak total read repairs



Riak connections



Fetcher disk caches



Summary

Buzzwords

Be amazed.

Doubt.

Evaluate.

Hardware

There is no such thing as
“too much RAM”

Scale

You'll need it.

Configuration Management

who's the master of puppet?

Monitoring

looks exciting even when things work.

Time.

Operational Stability beats Features
when it comes to Big **A Lot of Data.**

Thank you.

@geidies - seb@meltwater.com

<http://underthehood.meltwater.com/>

slides w/ notes on github.com/geidies/slides