Riak to the Rescue
Migrating Big Data

meltwater

meltwater

Big Data.

meltwater

Buzzwords.

Don't believe the Hype.

meltwater

Who am I?

Support
Development
SysAdmin
Managing Operations

Operations

8 Ops Engineers
4 Offices

Operations

650 physical
200 virtual
3 data centres

meltwater

Contact

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

meltwater

Migrating Big Data

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

meltwater

Meltwater

meltwater

Meltwater

News Monitoring

meltwater

Paper-Clip

Read News
Cut and Glue
Telefax

meltwater

Meltwater News

Crawl the Web
Match new Articles
Morning Report
Analytics UI

Products

PR

m|news
m|press

Marketing

m|buzz / engage
icerocket

meltwater

SaaS

Subscription model
24,000 clients


basho

meltwater

riak

- Open Source
- Dynamo Paper
- Erlang

meltwater

2.0

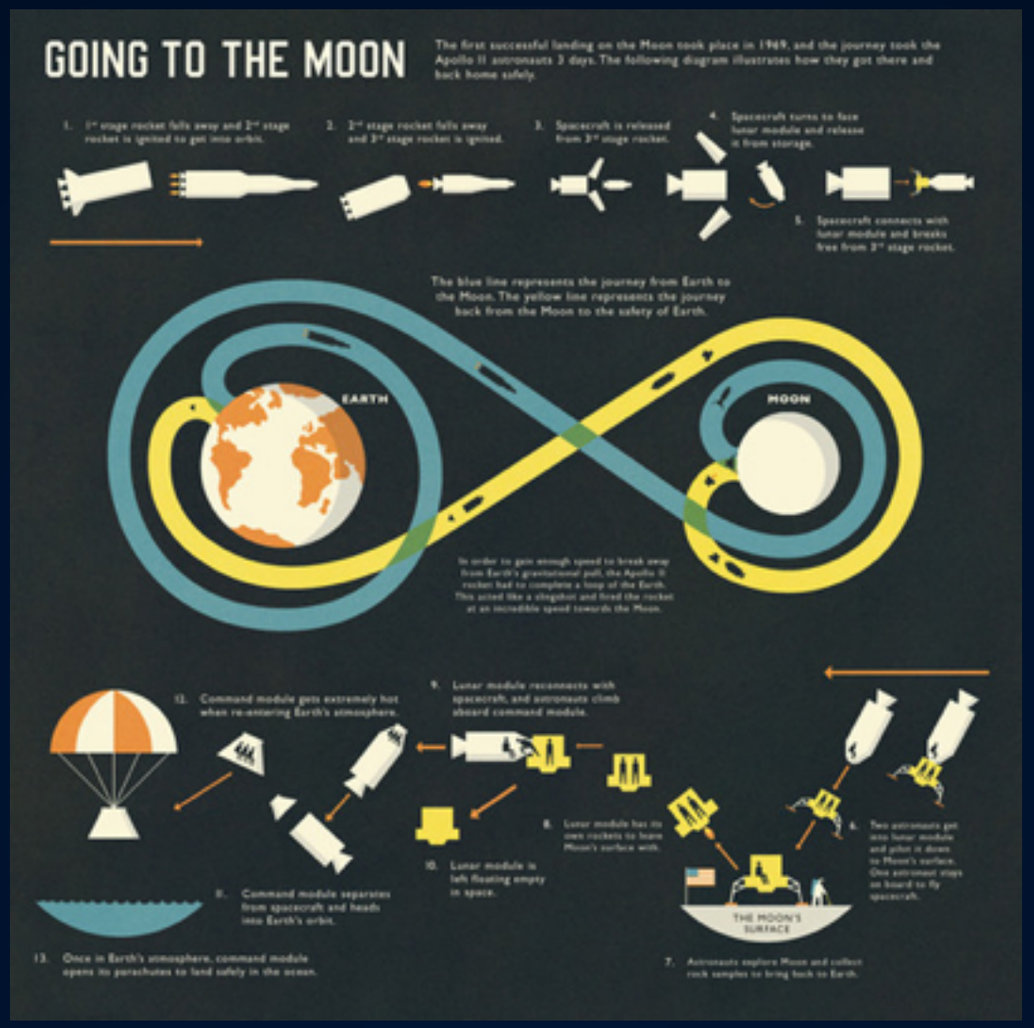
OMG, OMG!!

meltwater

thanks, basho.

meltwater

Meltwater Buzz



m|news 20 D/s - 8400 S/s

m|buzz 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 of foosball. You can find us at Rotherstraße 22 in Friedrichshain.

meltwater

TM

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

meltwater

Attention!

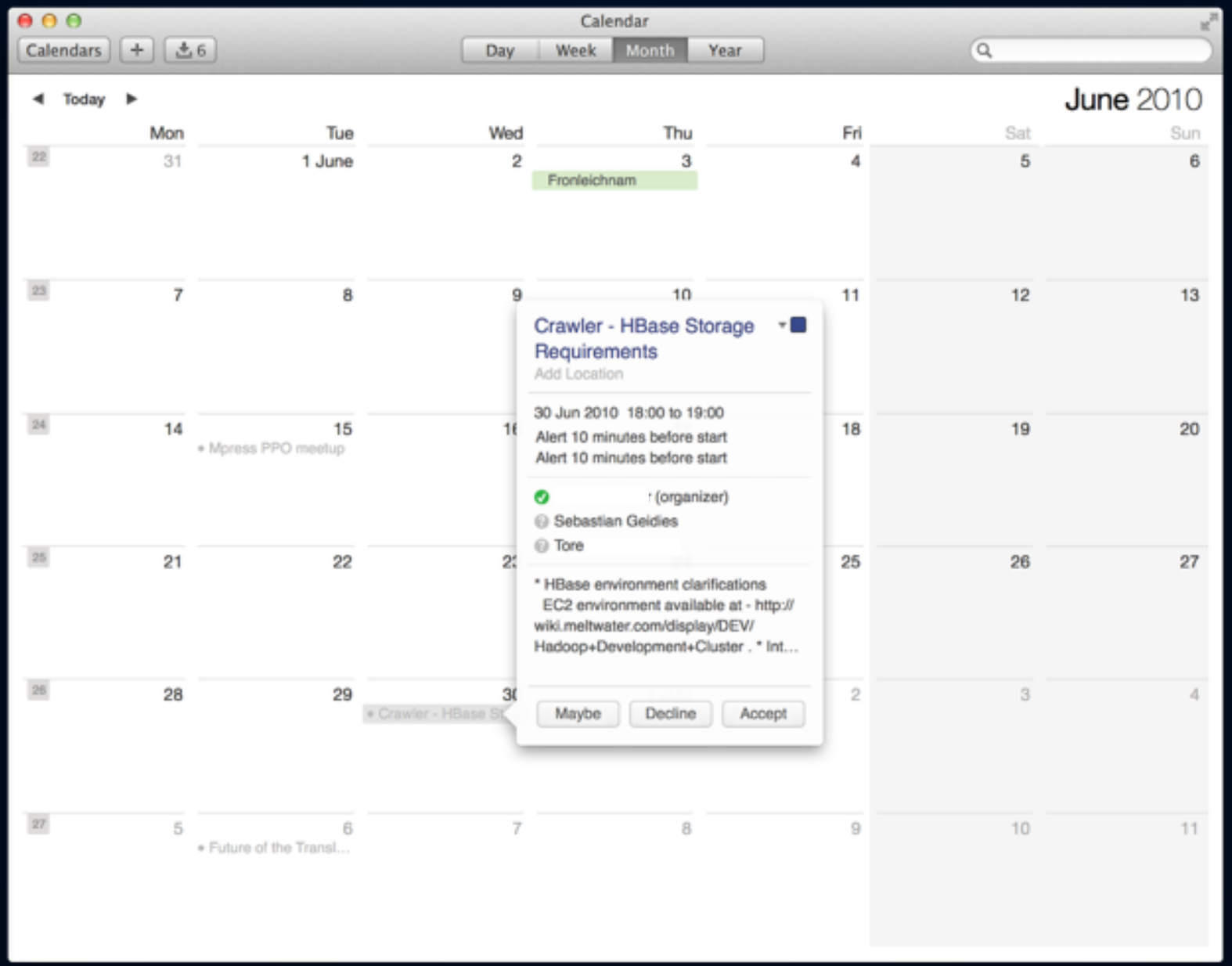
Your Use Case

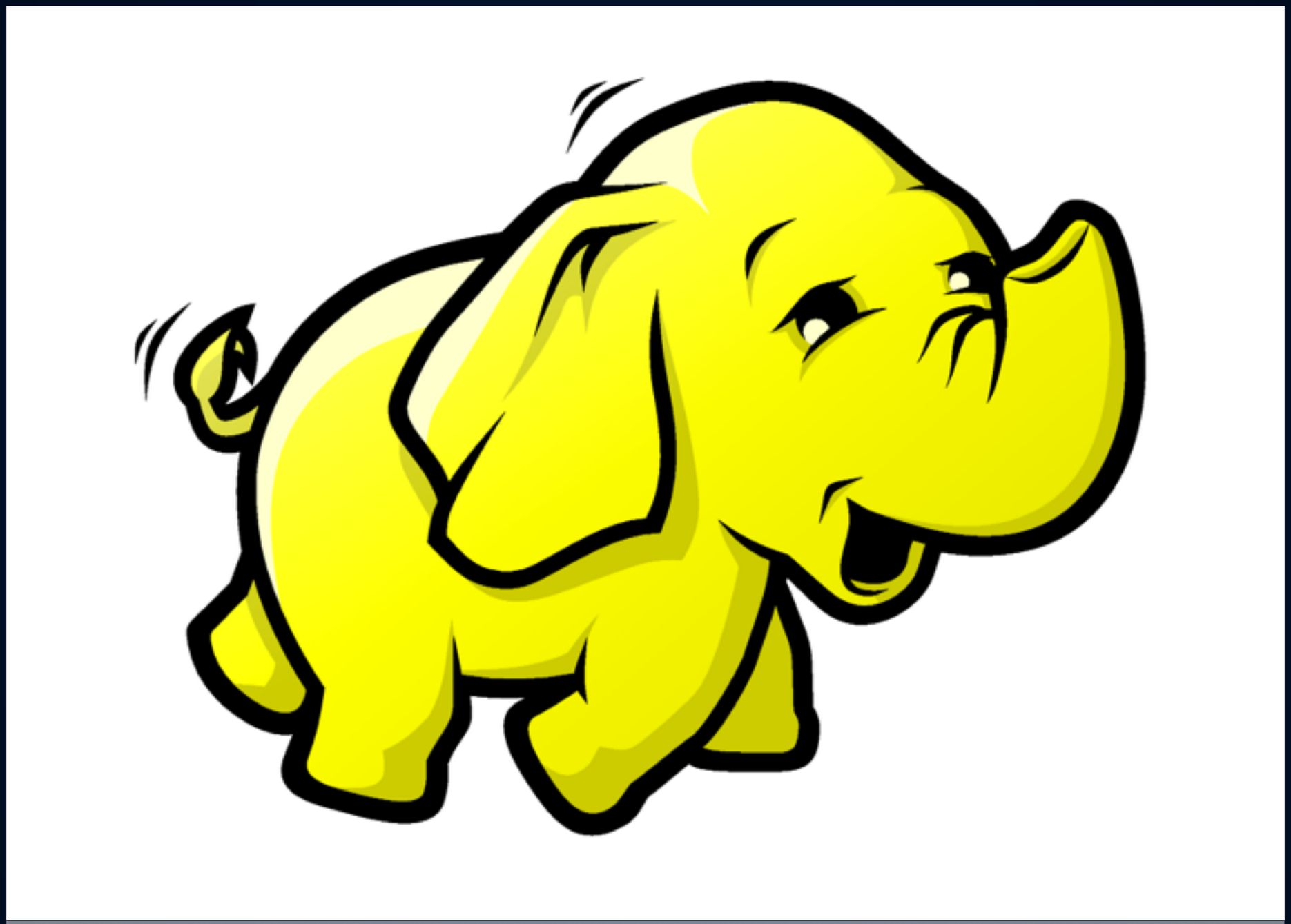
Research
Evaluate
Test

meltwater

m|buzz version 2

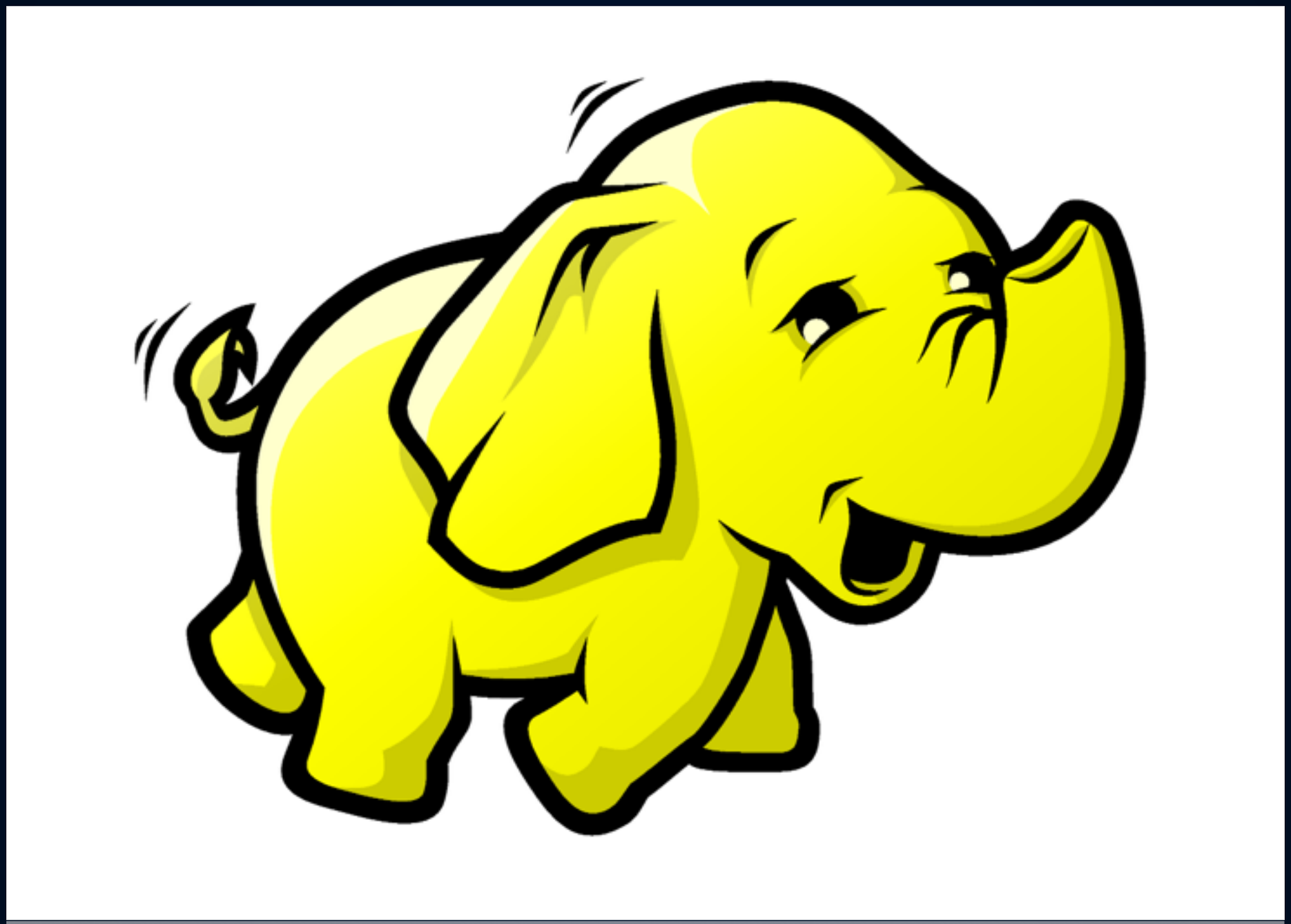
Scalability,
Features,
Buzzwords!


Calendar
Calendars
6
Day
Week
Month
Year
Today
June 2010
Mon
Tue
Wed
Thu
Fri
Sat
Sun
Fronleichnam
Mpress PPO meetup
Crawler - HBase Storage Requirements
Add Location
30 Jun 2010 18:00 to 19:00
Alert 10 minutes before start
Alert 10 minutes before start
(organizer)
Sebastian Geidies
Tore
* HBase environment clarifications
EC2 environment available at - http://wiki.meltwater.com/display/DEV/Hadoop+Development+Cluster . * Int…
Maybe
Decline
Accept
Future of the Transl…



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

– Jamie Zawinski



meltwater

Requirements

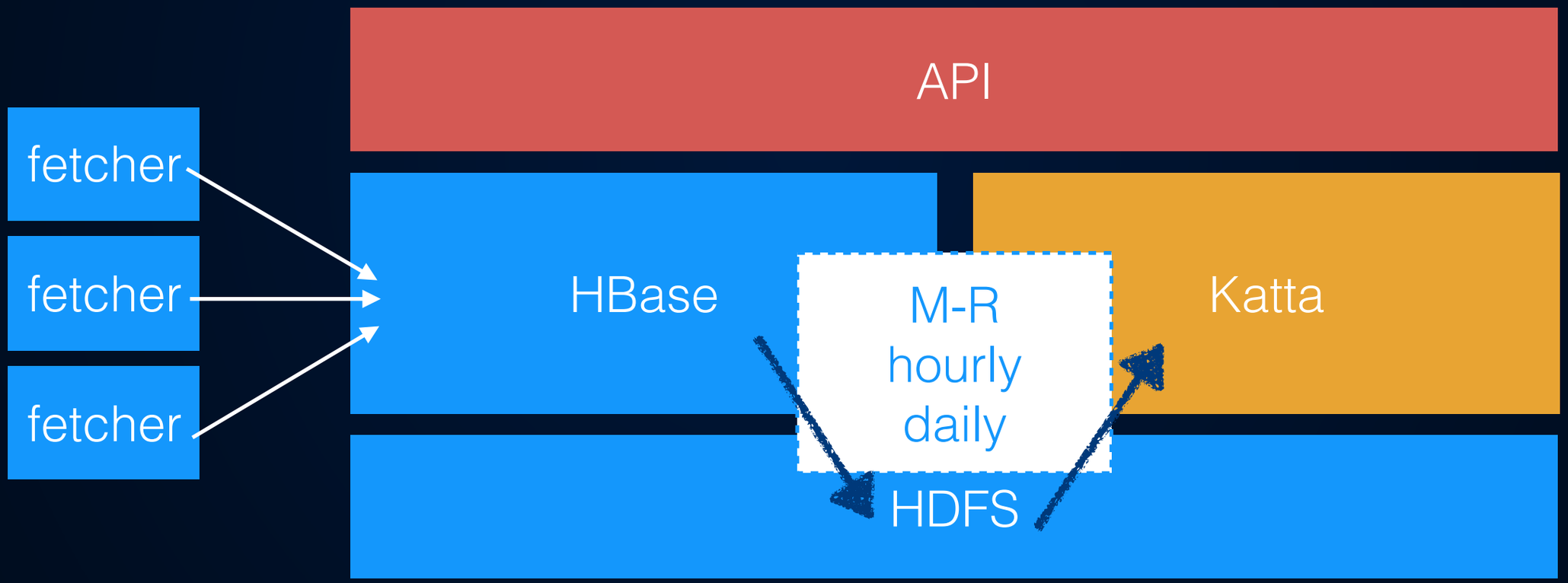
- Fail-Safety
- High Availability
- A Lot of Unstructured Data
- Near-Real-Time Indexing
- Time-Based Ordering instead of Relevancy

meltwater

m|buzz version 2

- Hadoop Ecosystem
- Apache Projects

m|buzz version 2



meltwater

It's a trap!

- buzzwords
- commodity hardware
- scale

meltwater

/opt/software

meltwater

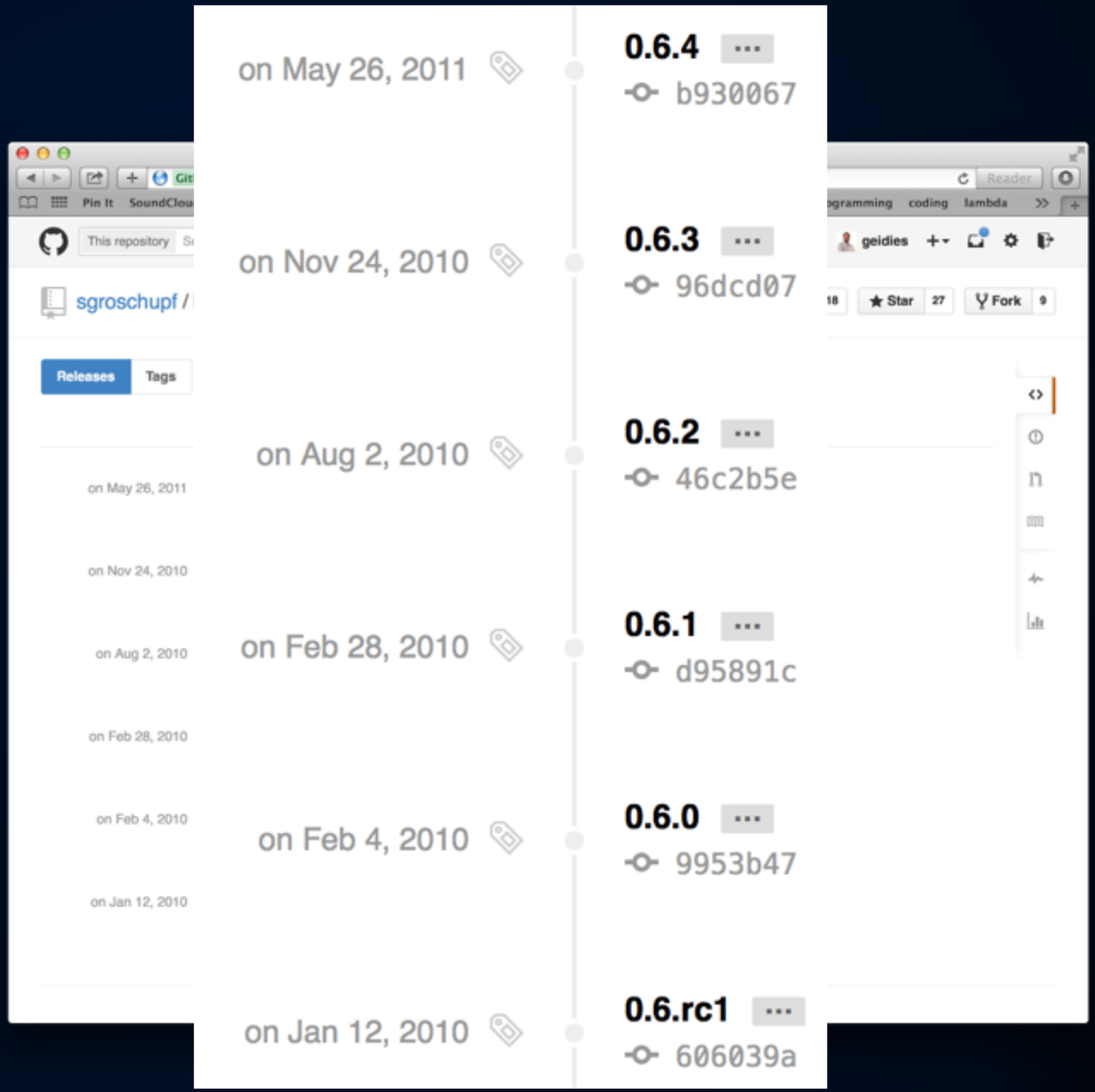
- Build upon lucene
- Master -> Worker -> Client
- communication through zookeeper
- multiple index copies
- copied from HDFS -> local disk

meltwater

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

meltwater

katta

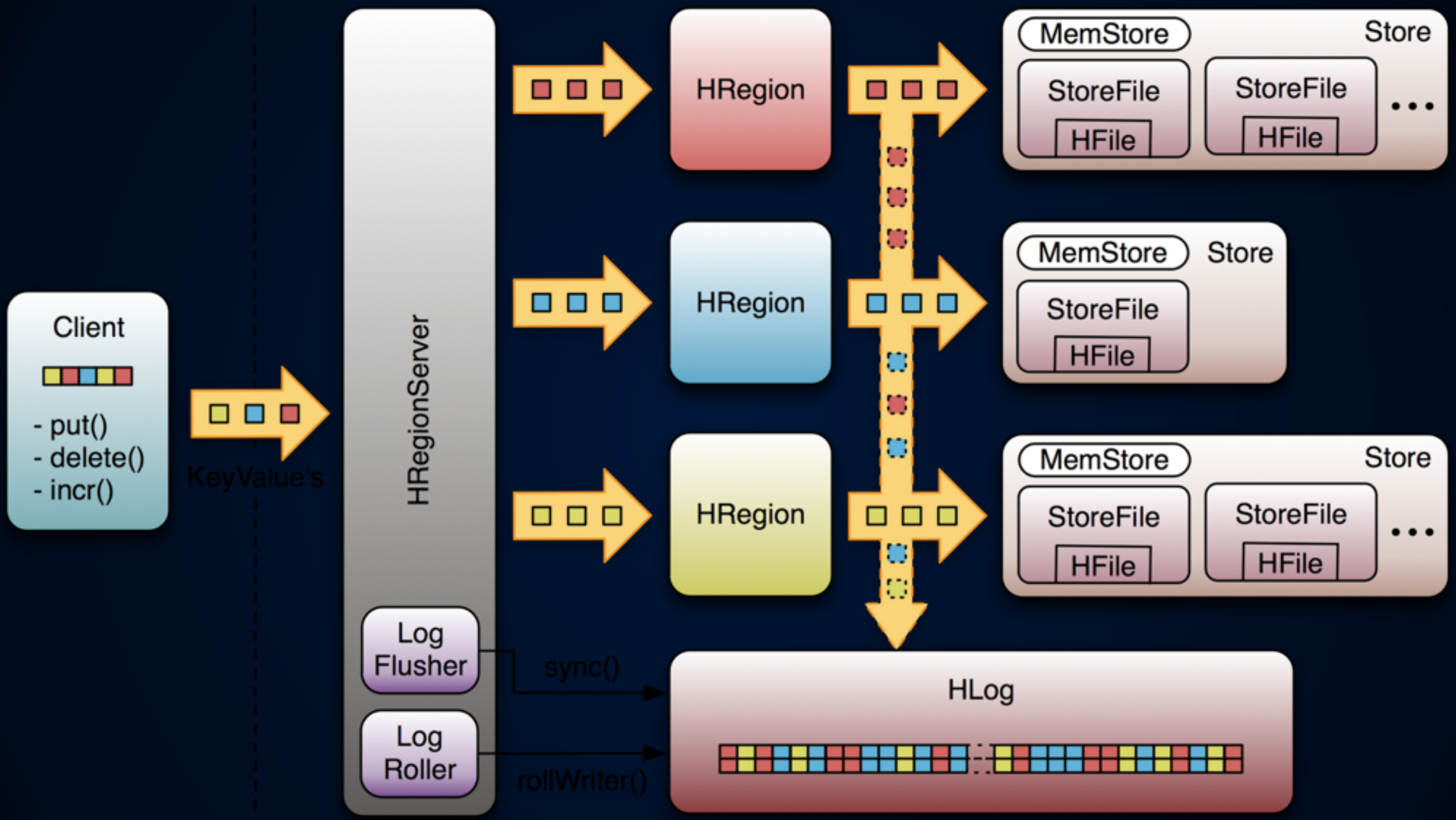


meltwater

meltwater

2025

meltwater


Client
- put()
- delete()
- incr()
KeyValue's
HRegionServer
Log Flusher
Log Roller
sync()
rollWriter()
HRegion
HRegion
HRegion
MemStore
Store
StoreFile
HFile
StoreFile
HFile
...
MemStore
Store
StoreFile
HFile
MemStore
Store
StoreFile
HFile
StoreFile
HFile
...
HLog

meltwater

split

keyspace

key a -> key c

key n -> key o

-ROOT-

.META.

meltwater

Fail-Safety

meltwater

Fail-Safety

Does NOT mean High Availability

Data on a Single Node

meltwater

Minutes.

55,000 posts / minute

meltwater

Funny Regions

Overlapping
Gaps
Negative Length

meltwater

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'}]}}

meltwater

HBase

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

meltwater

Good News!

NameNode never crashed.
Great.

meltwater

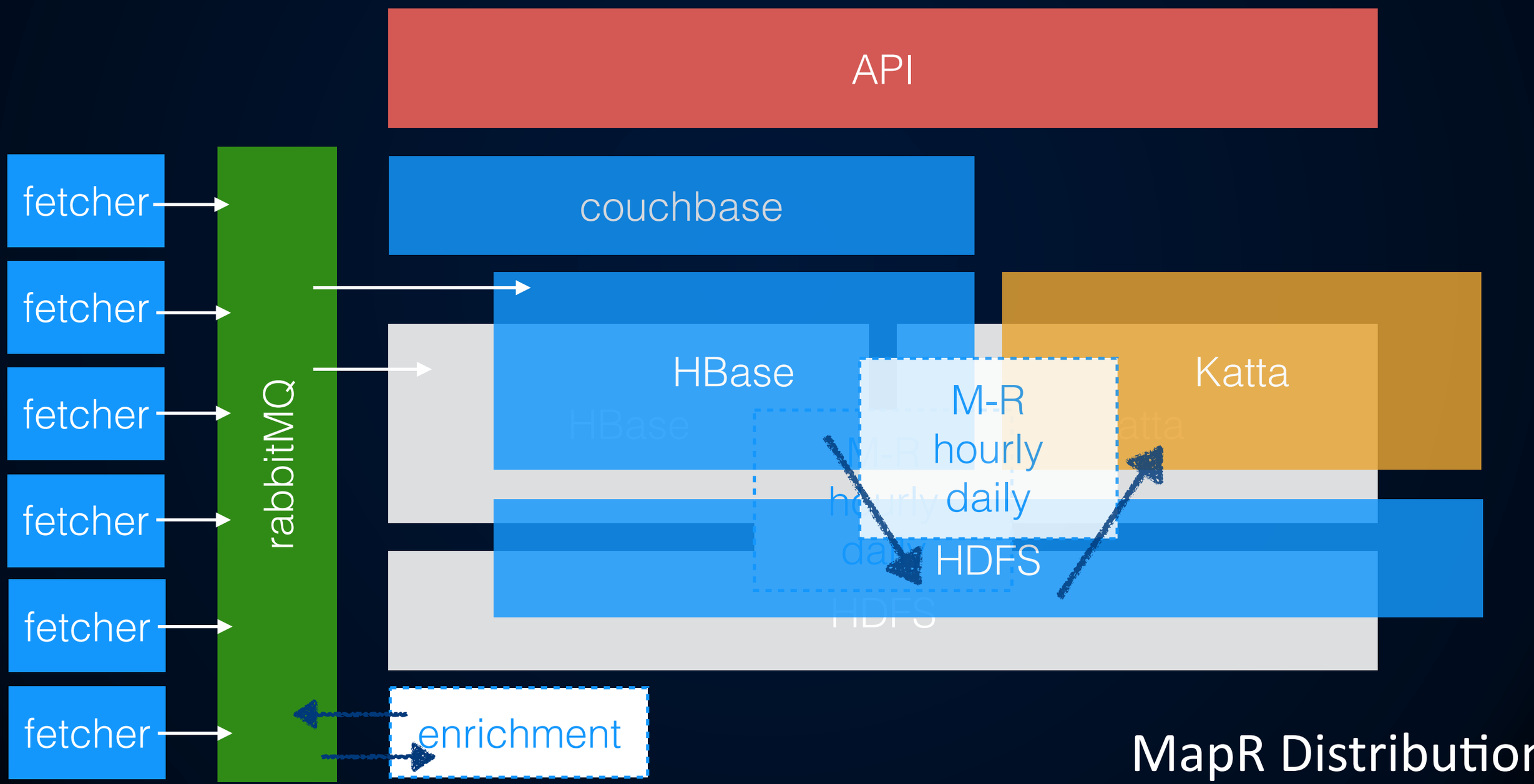
Changes...

…do you speak it?

meltwater

MAPR®

m|buzz version 2.5




Your logic makes the
puppy sad.

meltwater

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

m|buzz version 2.5

API

couchbase

fetcher

fetcher

fetcher

fetcher

fetcher

rabbitMQ

HBase

Katta

M-R hourly daily

HDFS

enrichment

MapR Distribution

meltwater


membase


CouchBase

meltwater

First Read Wins

Parallel Reads:
couchbase
vanilla HBase
MapR HBase

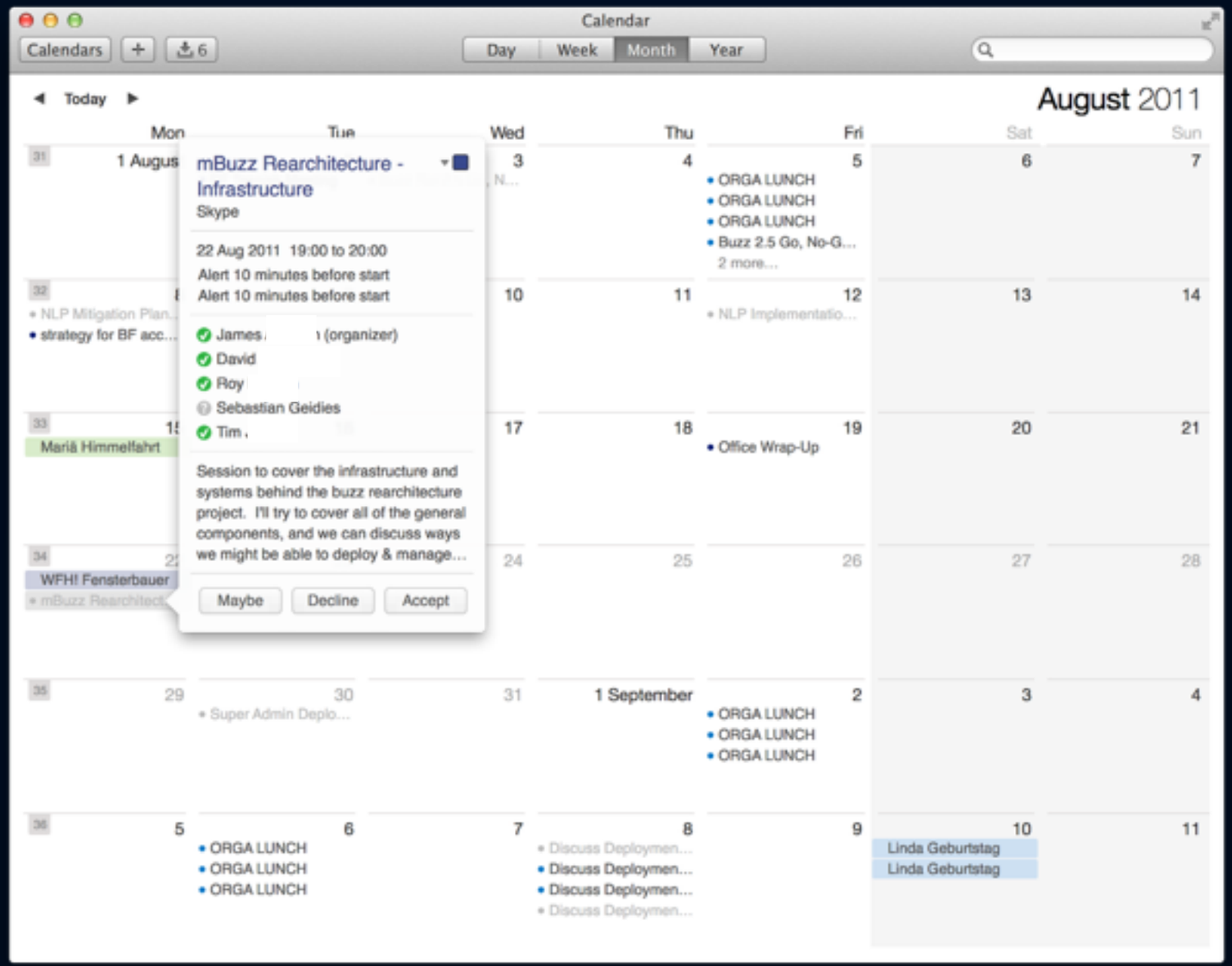
meltwater

couchbase scales!

…to four weeks of data.
2.2B entries
TTL

meltwater

Are we there yet?


Calendar
Calendars
6
Day
Week
Month
Year
Today
August 2011
Mon
Tue
Wed
Thu
Fri
Sat
Sun
mBuzz Rearchitecture - Infrastructure
Skype
22 Aug 2011 19:00 to 20:00
Alert 10 minutes before start
Alert 10 minutes before start
James (organizer)
David
Roy
Sebastian Geidies
Tim
Session to cover the infrastructure and systems behind the buzz rearchitecture project. I'll try to cover all of the general components, and we can discuss ways we might be able to deploy & manage...
Maybe
Decline
Accept
ORGA LUNCH
Buzz 2.5 Go, No-G...
2 more...
NLP Implementatio...
Office Wrap-Up
Mariä Himmelfahrt
WFH! Fensterbauer
Super Admin Deplo...
1 September
Discuss Deploymen...
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		

meltwater

Requirements

✓ High Availability

✓ Data Safety

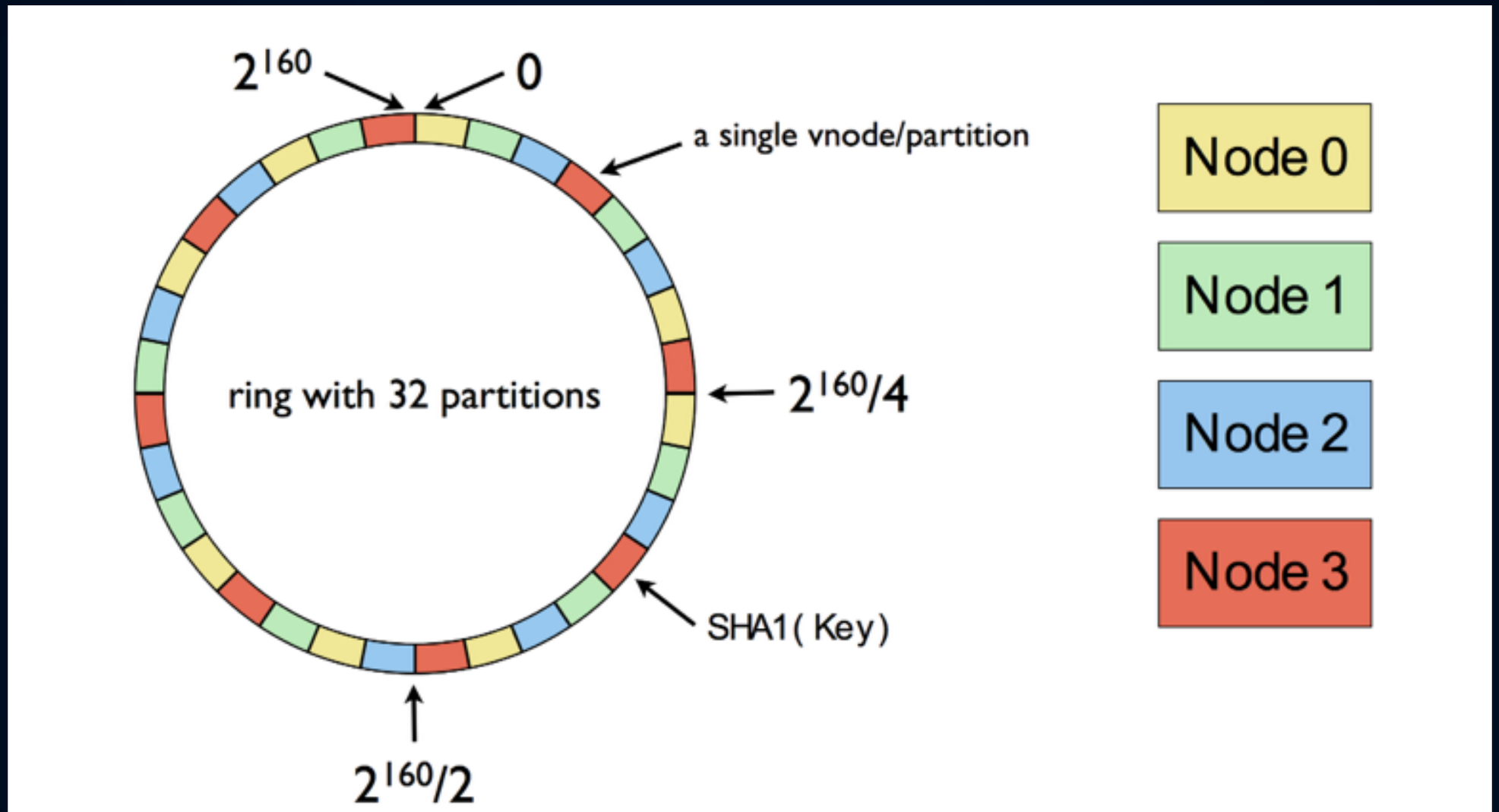
✓ Scalability

? Range Scans or TTL to limit data

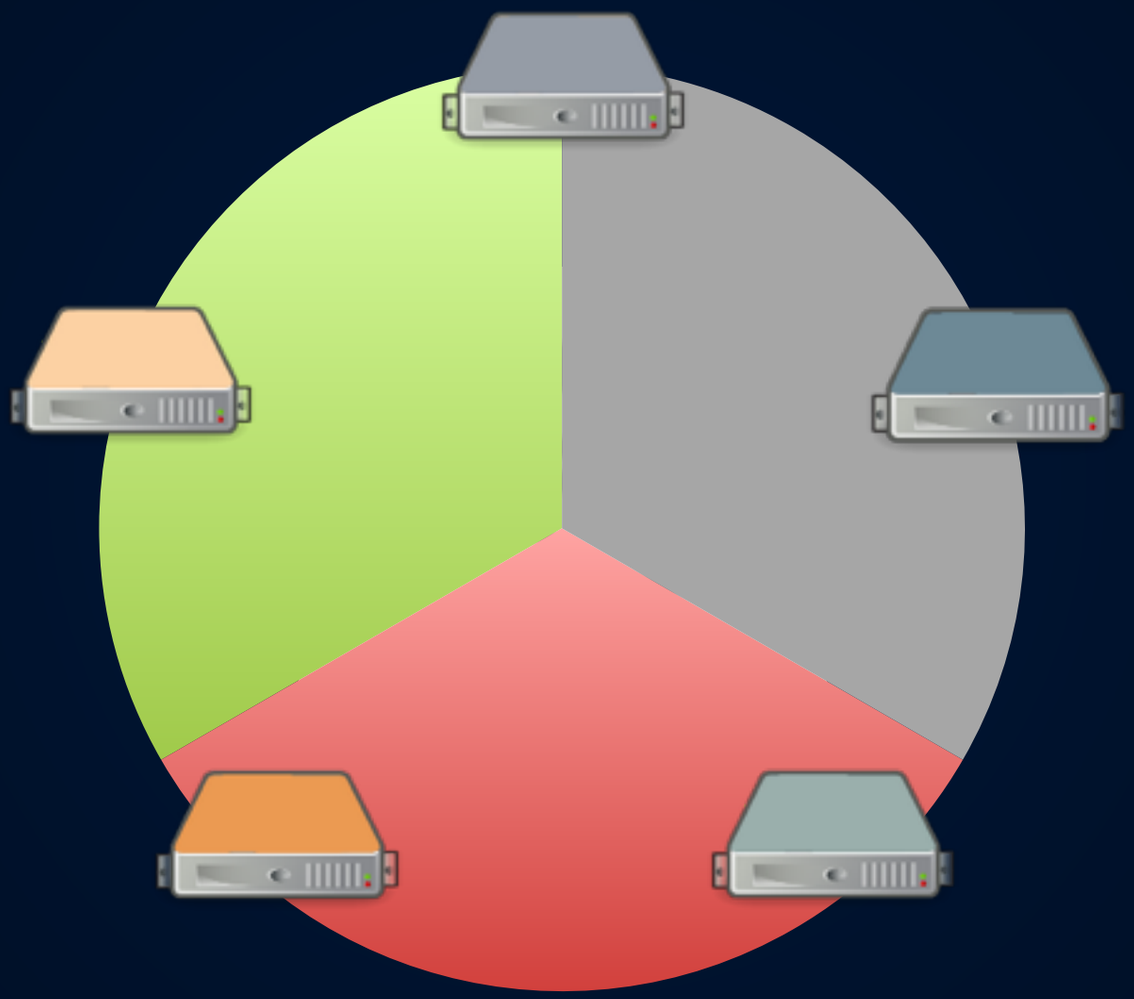
meltwater

riak

Key-Value model
Objects in
Buckets


2^160
0
a single vnode/partition
ring with 32 partitions
2^160/4
SHA1(Key)
2^160/2
Node 0
Node 1
Node 2
Node 3

meltwater

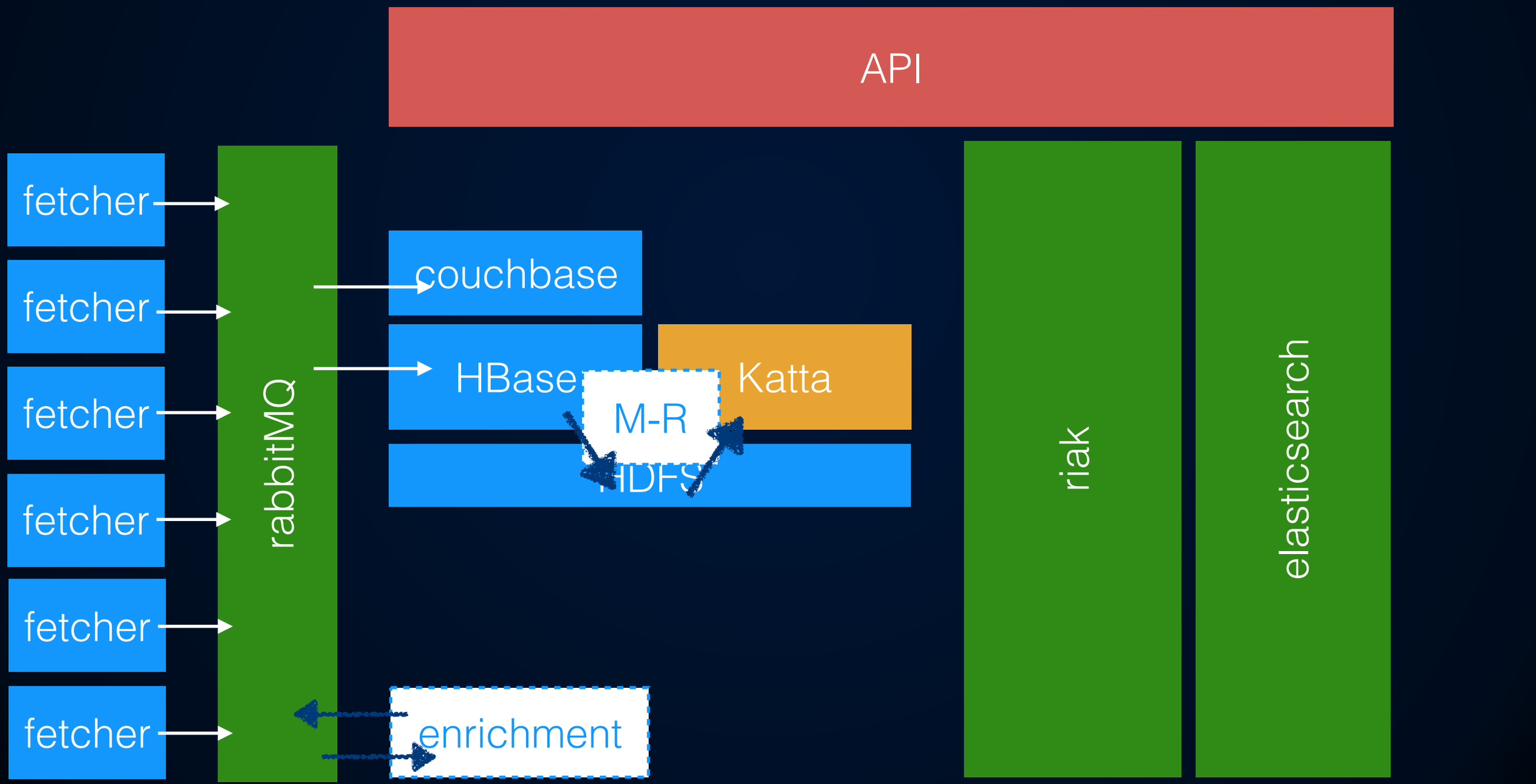


"While there are mechanisms such as Vector Clocks to help deal with these issues, if your application requires the kind of strong consistency found in ACID systems, Riak may not be a good fit."

– riak documentation

meltwater

m|buzz version 2.6



meltwater

Commodity Hardware

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

meltwater

Configuration

- levelDB
- erlang VM
- Map-Reduce

meltwater

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

meltwater

Numbers

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

Hey, wait.

A good three weeks?

meltwater

Let’s do it.

parallel reads
gather numbers
stability
speed

meltwater

riak is slow.

but consistent,
and massively parallel.

meltwater

~~riak is slow.~~

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

meltwater

stability over speed.

meltwater

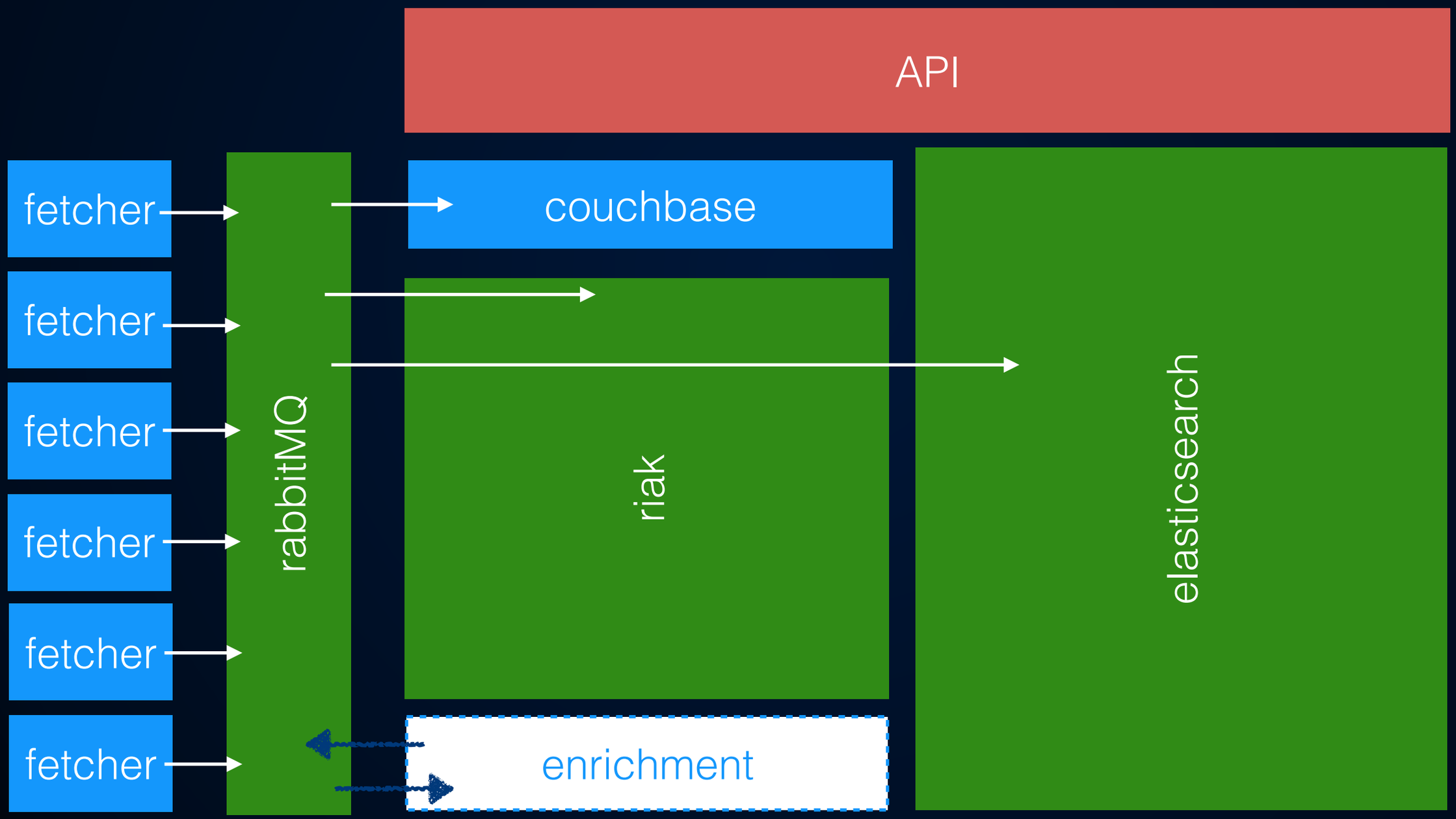
stability

- availability during
  - node failures
  - upgrades
  - configuration updates

meltwater

Search

m|buzz version 3

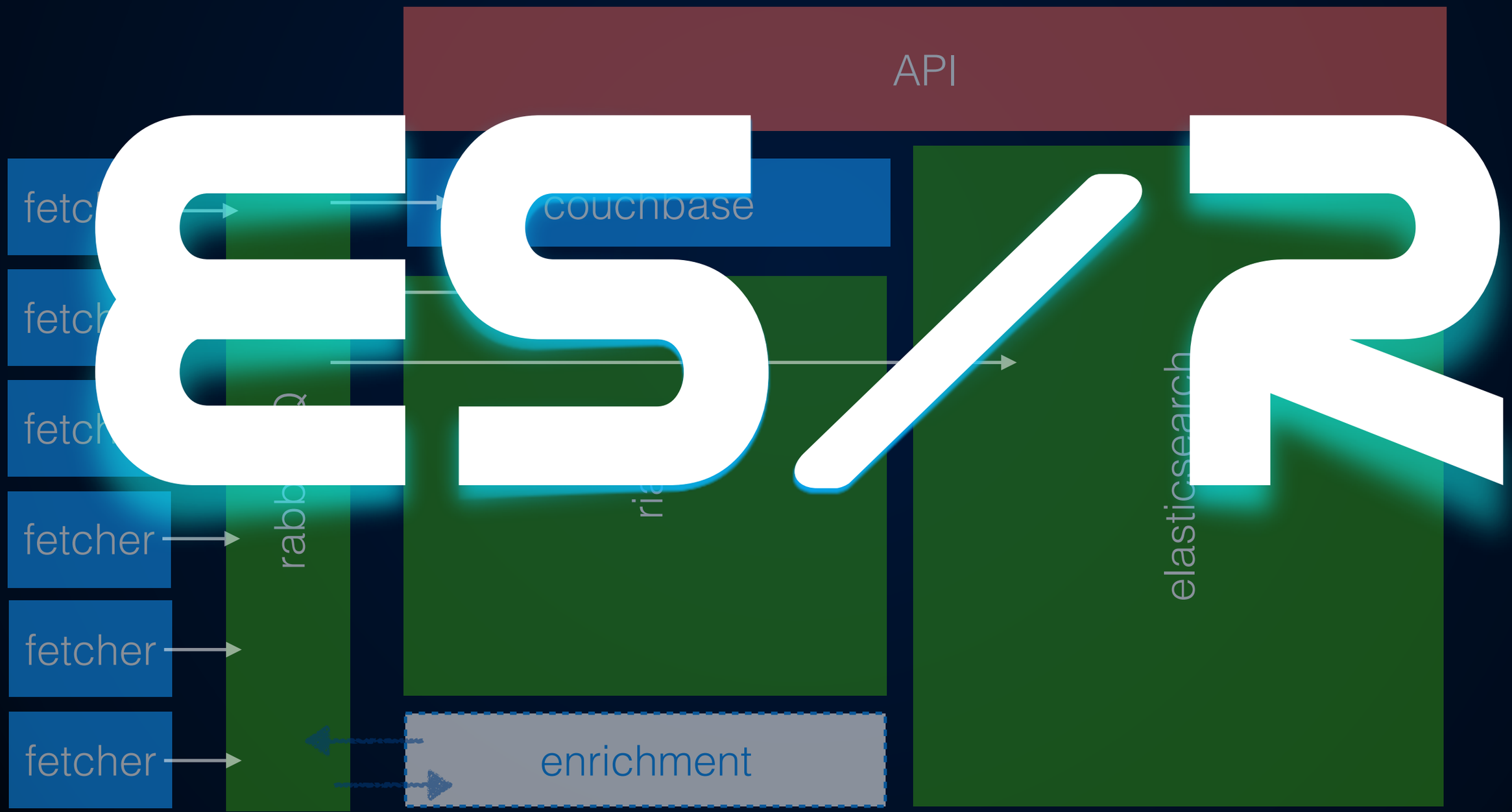


meltwater

Naming Things

meltwater

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. :)

meltwater

Still live

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

meltwater

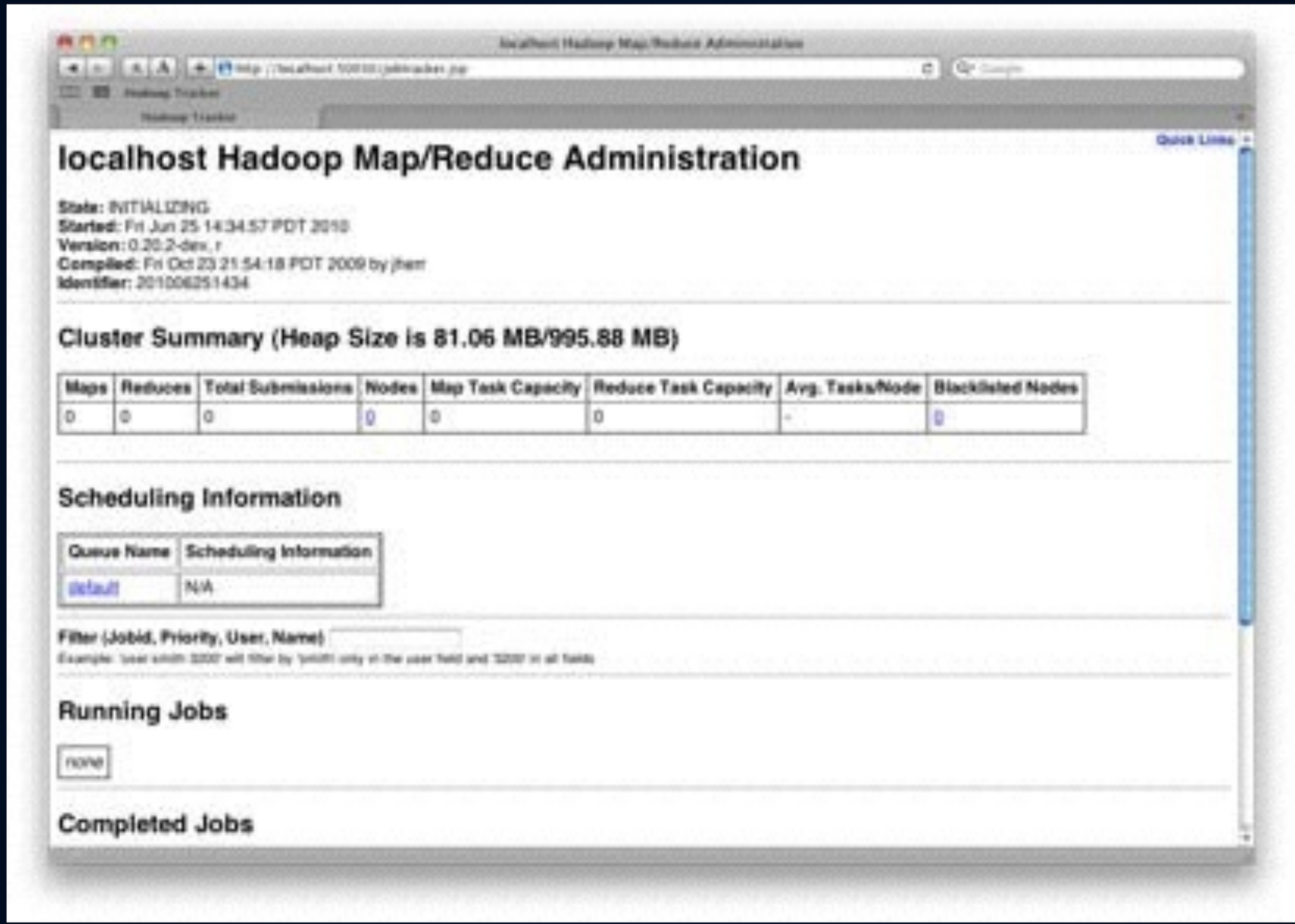
Monitoring

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

meltwater

Dashboards

And APIs.



localhost Hadoop Map/Reduce Administration

http://localhost:50030/jobtracker.jsp

Hadoop Tracker

Hadoop Tracker

Quick Links

localhost Hadoop Map/Reduce Administration

State: INITIALIZING
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 jhen
Identifier: 201006251434

Cluster Summary (Heap Size is 81.06 MB/995.88 MB)

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

Scheduling Information

Queue Name	Scheduling Information
default	N/A

Filter (Jobid, Priority, User, Name)

Example: 'user smith 3200' will filter by 'smith' only in the user field and '3200' in all fields

Running Jobs

none

Completed Jobs

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

Local logs, Thread Dump, Log Level, Debug dump

Attributes

Attribute Name	Value	Descrip
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 compile
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 compil
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-ac40-76d5ffacb455	Unique identifier generated for ea
Load average	0	Average number of regions per re
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 ser
Coprocessors	[]	Coprocessors currently loaded loa
HMaster Start Time	Tue Feb 21 19:18:24 EST 2012	Date stamp of when this HMaster
HMaster Active Time	Tue Feb 21 19:18:24 EST 2012	Date stamp of when this HMaster

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

https rabbit-ops.melt.no/#/queues/%2F/pipeline

Pin It | SoundCloud Download | MNews stack | Passpack It! | DuckDuckGo | Wikipedia | News | Beliebt | 0to255 | programming | coding | lambda

RabbitMQ

User: Administrator
RabbitMQ 3.1.5, Erlang R14B04
Log out

Overview | Connections | Channels | Exchanges | Queues | Admin

Virtual host: /

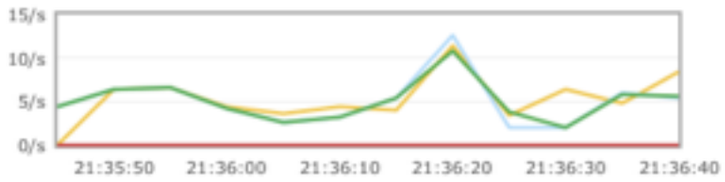
Queue pipeline

Overview

Queued messages (chart: last minute) (?)

Ready	0 msg
Unacknowledged	8 msg
Total	8 msg

Message rates (chart: last minute) (?)



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

Details

Parameters	x-message-ttl: 600000	Status	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

Couchbase Console (1.8.1)

https couchbase.ops.melt.no/index.html#sec=analytics&statsHostname=10.31.0.192:8091

Reader

Pin It | SoundCloud Download | MNews stack | Passpack It! | DuckDuckGo | Wikipedia | News | Beliebt | 0to255 | programming | coding | lambda | gmail | Clients From Hell | gotAPI | Rails API | R-Doc Library

Couchbase

About · Sign Out

MONITOR

- Cluster Overview
- Data Buckets
- Server Nodes
- Log

MANAGE

- Data Buckets
- Server Nodes
- Settings

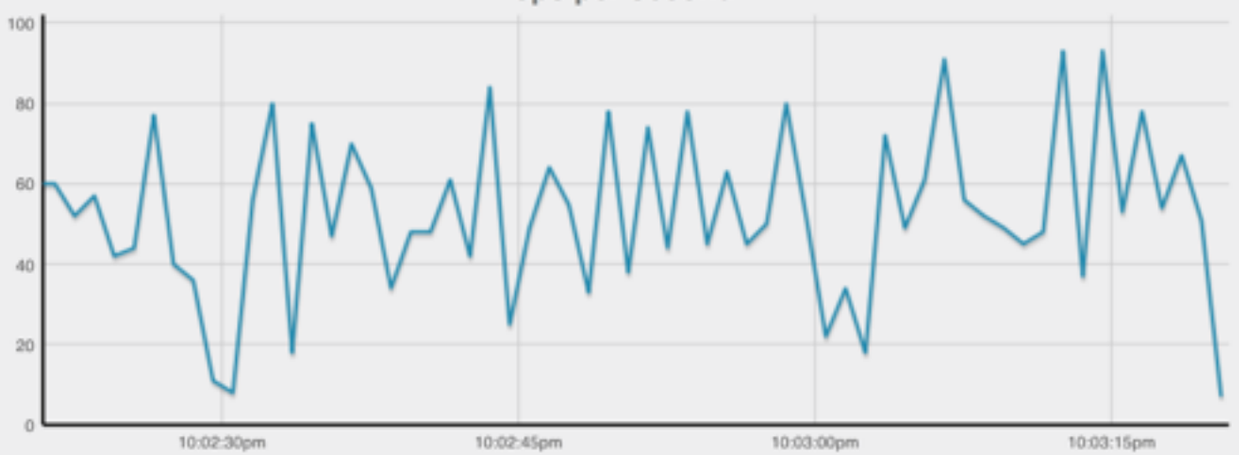
SUPPORT

- Documentation
- Support Forums

DATA BUCKETS: default on 10.31.0.192

Minute | Hour | Day | Week | Month | Year — Last 1 minute

ops per second



SERVER RESOURCES

151K swap usage | 6.71G free RAM | 1.25 CPU utilization % | 39 connections

SUMMARY

7 ops per second | 0 cache miss ratio | 0 creates per sec. | 0 updates per sec.

0 disk reads per sec. | 0 temp OOM per sec. | 0 gets per sec. | 7 sets per sec.

0 CAS ops per sec. | 0 deletes per sec. | 23.1M items | 3.77K disk write queue

21.5G memory used | 22.8G high water mark | 18.2G low water mark | 0 disk update time

VBUCKET RESOURCES

ACTIVE | REPLICA | PENDING | TOTAL


riak control
basho
Snapshot
Cluster
Ring
Objects
Map/Reduce
Graphs
Logs
Support
Ring View
Filter
All Owners
Prev
1
Next
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

meltwater

necessary but not sufficient

dashboard
API
fool-safe performance configuration
good documentation

meltwater

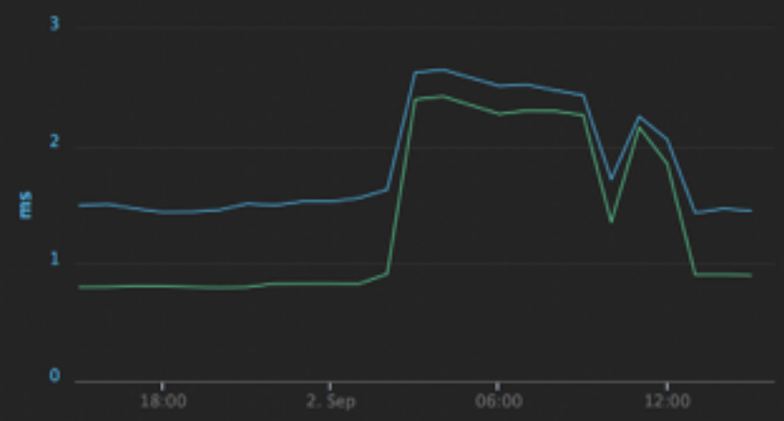
Dashboards | Buzz

SEP 01, 14 02:50 PM - SEP 02, 14 03:50 PM

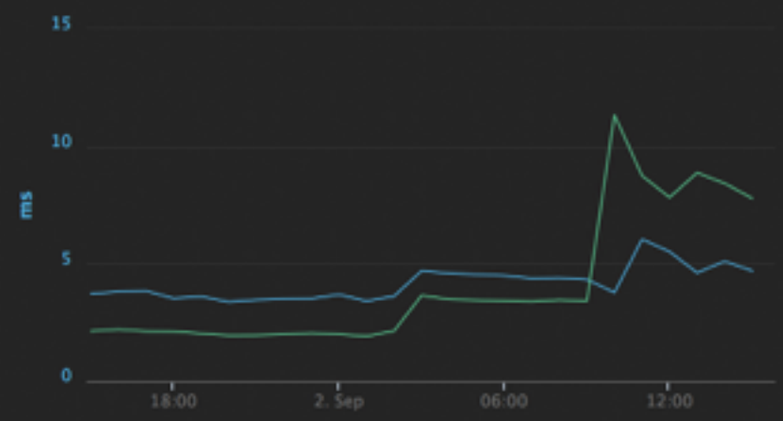
UTC

Riak GET/PUT count

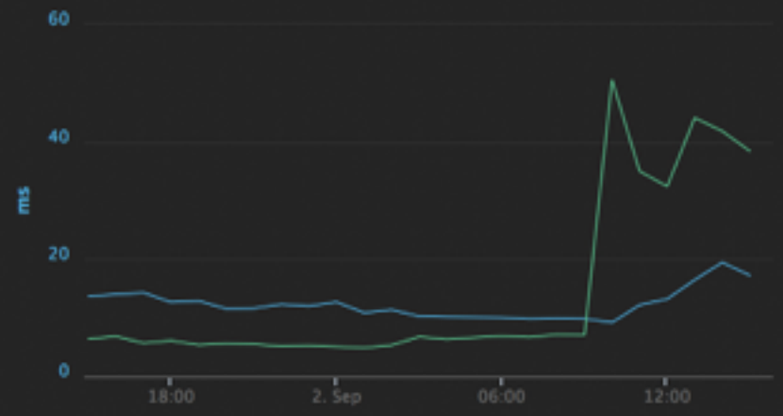
Riak GET/PUT latency (median)



Riak GET/PUT latency (mean)

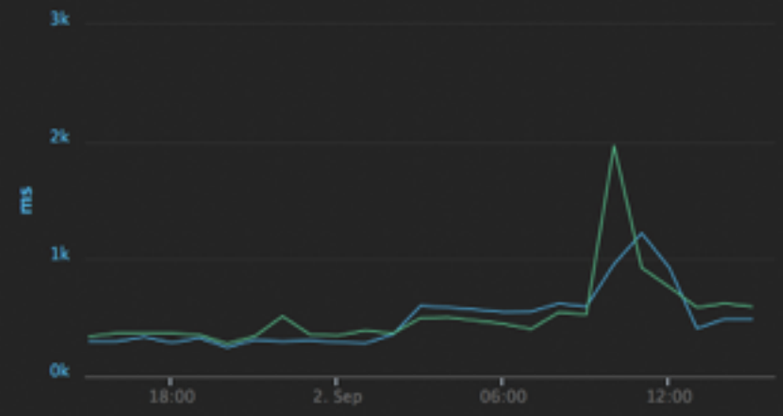


Riak GET/PUT latency (95th)

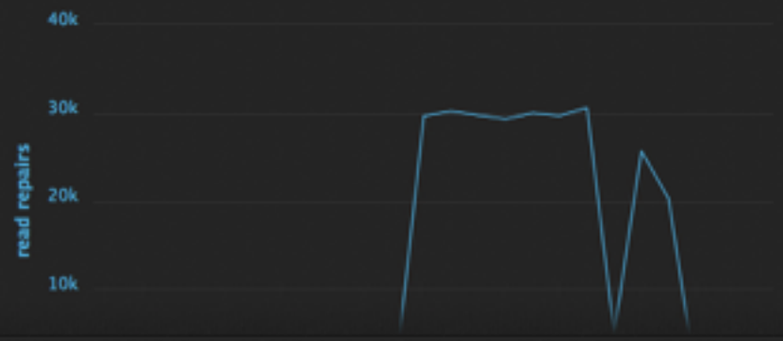


Riak GET/PUT latency (99th)

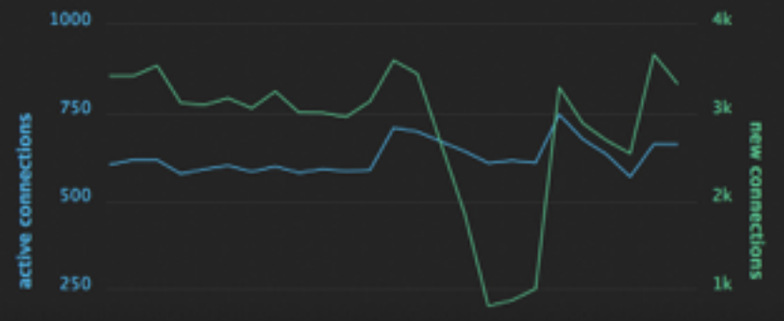
Riak GET/PUT latency (100th)



Riak total read repairs



Riak connections



Fetcher disk caches

THERE IS NO DATA FOR THIS TIME INTERVAL.

librato™

meltwater

Summary

meltwater

Buzzwords

Be amazed.
Doubt.
Evaluate.

meltwater

Hardware

There is no such thing as
“too much RAM”

meltwater

Scale

You'll need it.

meltwater

Configuration Management

who's the master of puppet?

meltwater

Monitoring

looks exciting even when things work.

meltwater

Time.

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

meltwater

Thank you.

@geidies - seb@meltwater.com

http://underthehood.meltwater.com/

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