Data driven applications

based on Apache projects

Image by Studio Tempura - http://www.flickr.com/photos/zero101/3335373387

Isabel Drost

Daytime:

Co-Founder Berlin Buzzwords.
Software developer at Nokia Gate 5 GmbH.

Nighttime:

Co-Founder Apache Mahout.
Member Apache Software Foundation.
Founder of Berlin Hadoop Get Together.

Hello GoTo Con visitors.

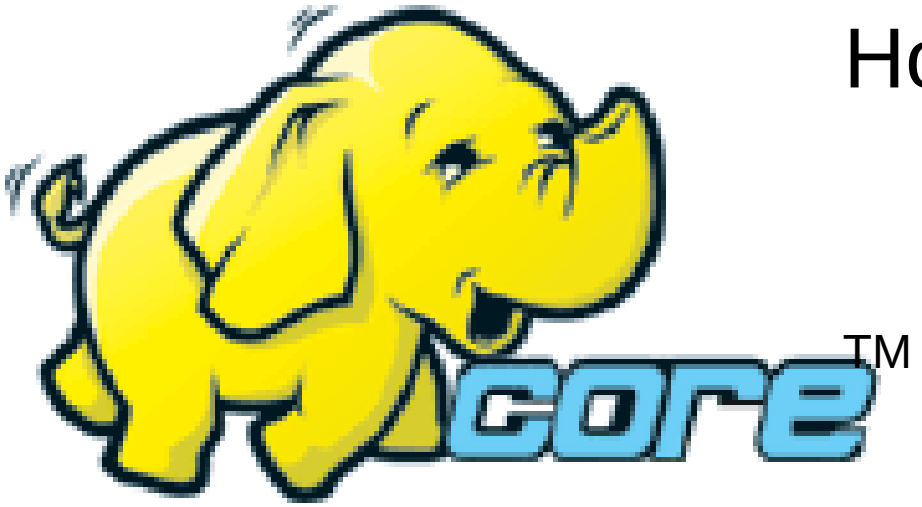
Apache Hadoop

Know it?

Use it?

How many nodes?

Contributed patches?



Apache Lucene/ Solr

Apache Mahout


amazon
web services™
AWS
Products
Developers
Products & Services
Amazon Elastic MapReduce

Amazon EMR


By: mondays child - http://www.flickr.com/photos/normanlowery/4729806370/

Agenda

- Introducing example
- Data driven design.
- Types of data.
- Toolset.

A highly representative example

Image thanks to Philip Kaden and Berlin Buzzwords
http://flickr.com/photomicc
http://flickr.com/berlinbuzzwords

Image by frank3.0: http://www.flickr.com/photos/frank3/6053973411/


WALL
PEOPLE
РОССИЯ
Image by Zaprittsky: http://www.flickr.com/photos/zaprittsky/4520788183/

Ways to uncover requirements?

BDUF

(big design up front)

On-site customer

(Product owner in Scrum)

Image by tö: http://www.flickr.com/photos/2ni/5814650182

Can we do better?

(Switch the oracle against metrics.)

Data-based approach

(“Get out of the building”)

Image by Leo Reynolds:
http://www.flickr.com/photos/lwr/3552240620

Learn from past users' behaviour

Users seldom give explicit feedback.

Take a snapshot of user interactions.

Analyze what they do.

Derive needs for new functionality from that.

Your search did not match any documents...

Searching for:

date when none is indexed.

project pig instead of the animal.

typing AMS but mean Amsterdam.

typing confrence but mean conference.

Discover content without effort by …

Showing content that I might like based on:

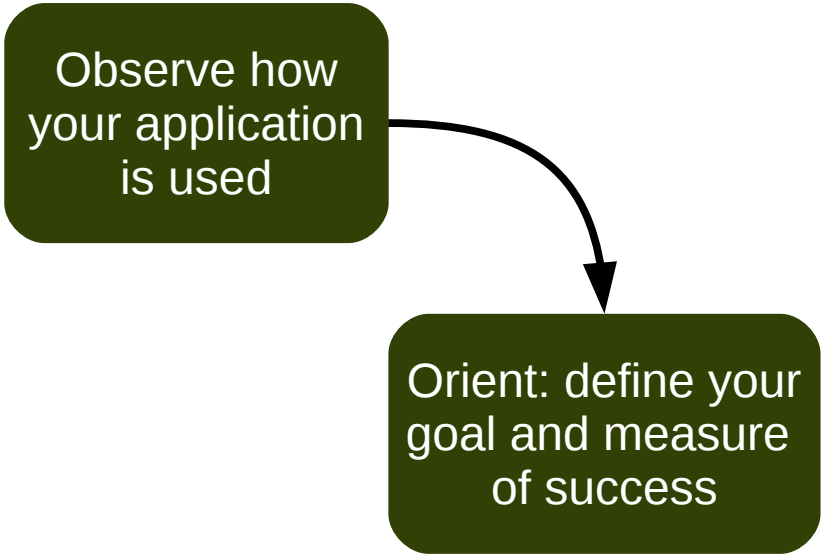
my personal interests and tags.

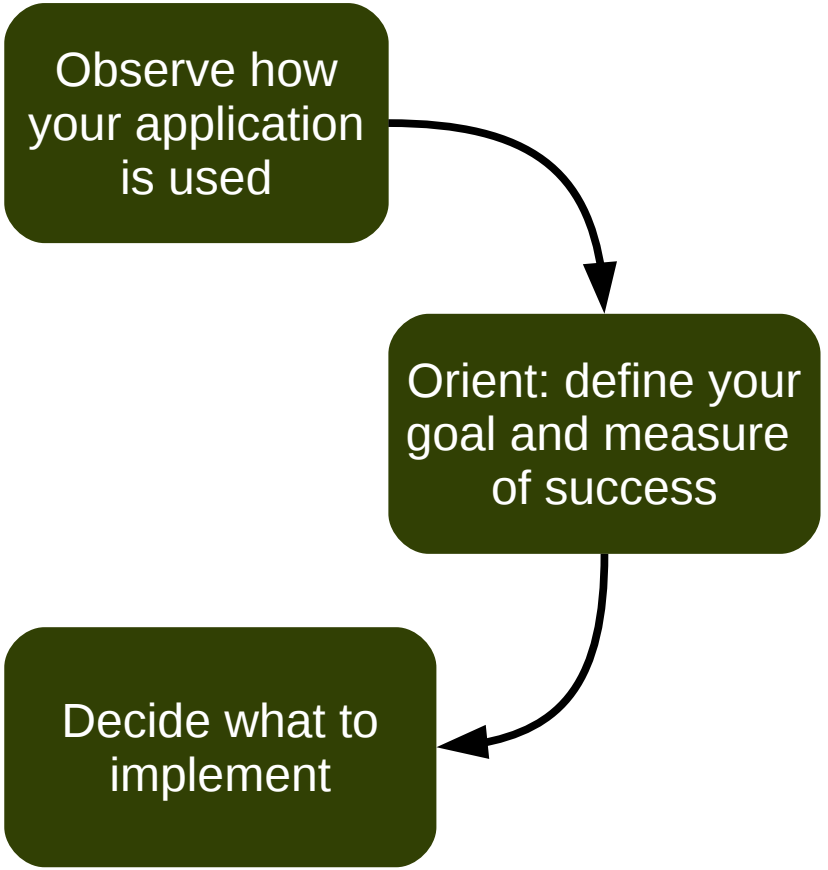
talks I already indicated as “want to attend.

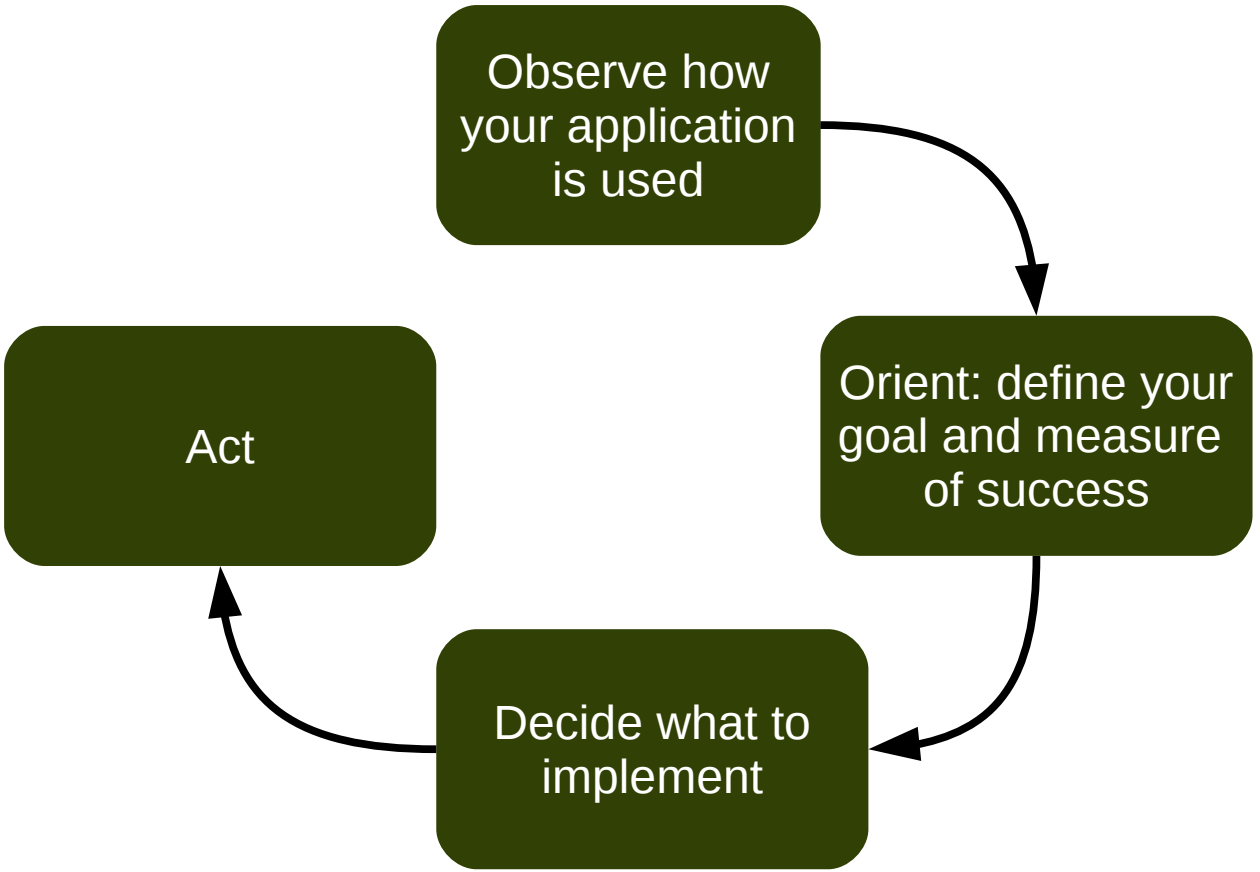
abstracts and speaker info pages I read.

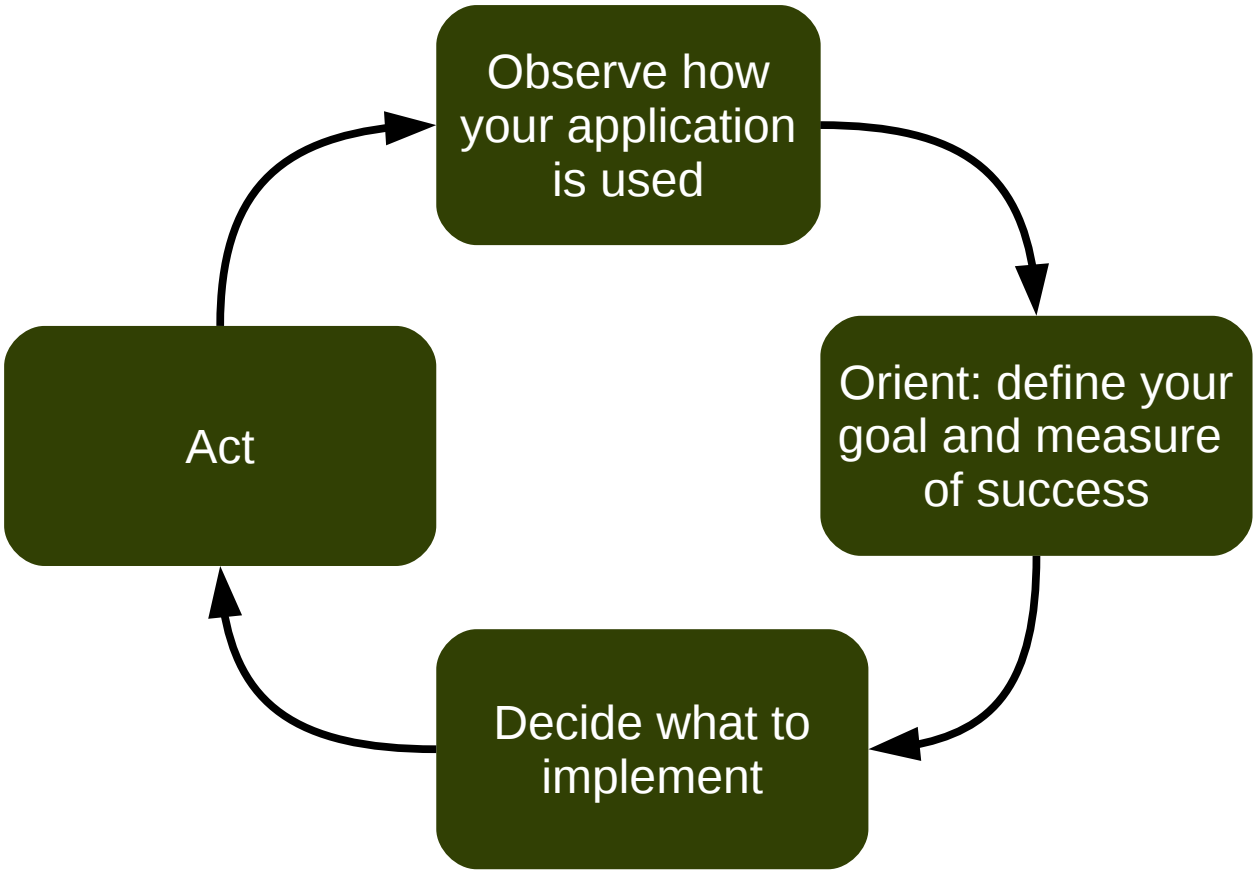
my prior search history and twitter messages.

Observe how your application is used


Observe how your application is used
Orient: define your goal and measure of success


Observe how your application is used
Orient: define your goal and measure of success
Decide what to implement


Observe how your application is used
Orient: define your goal and measure of success
Decide what to implement
Act



OODA by USAF strategis John Boyd

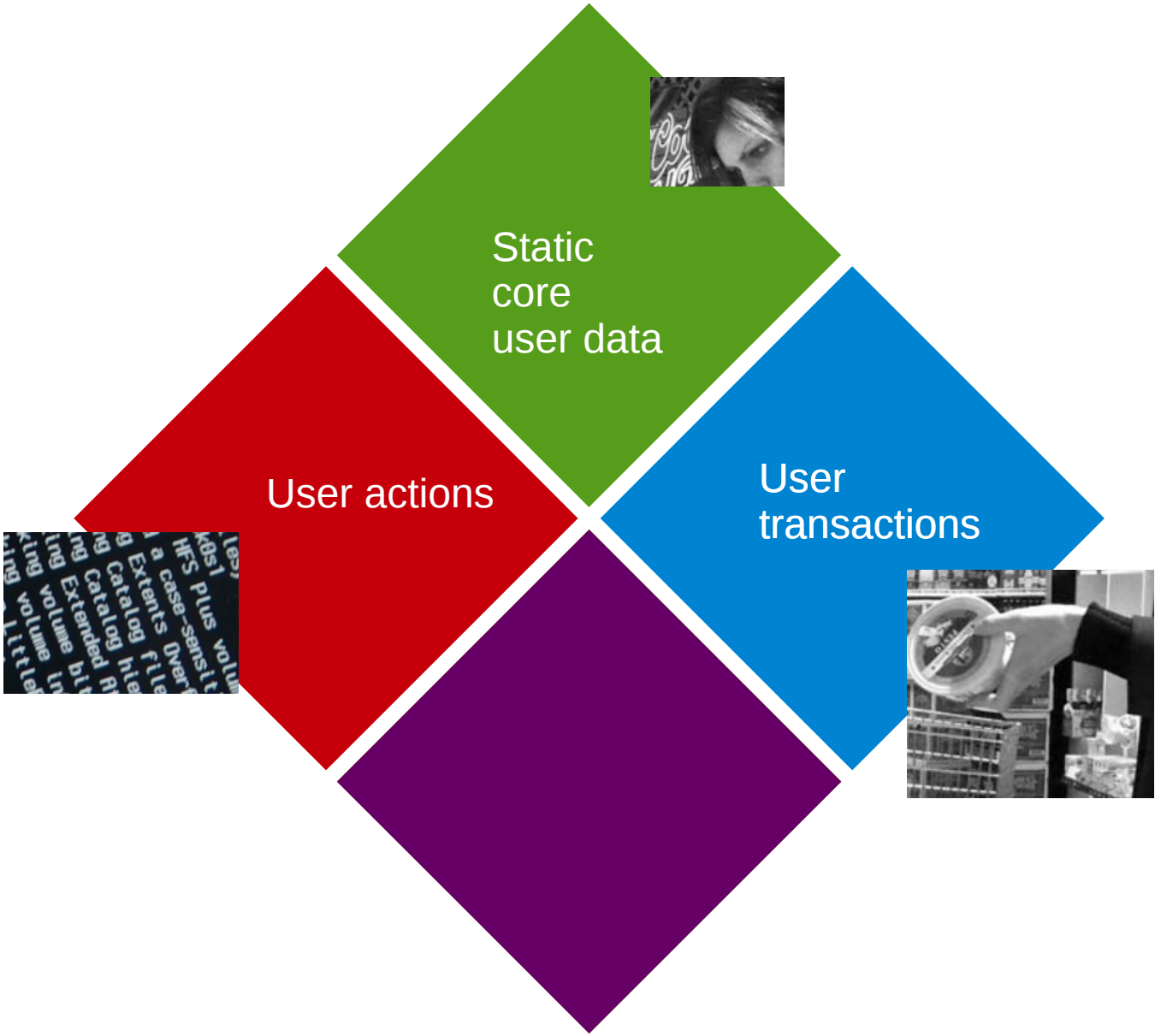
By Lab2112, http://www.flickr.com/photos/lab2112/462388595/

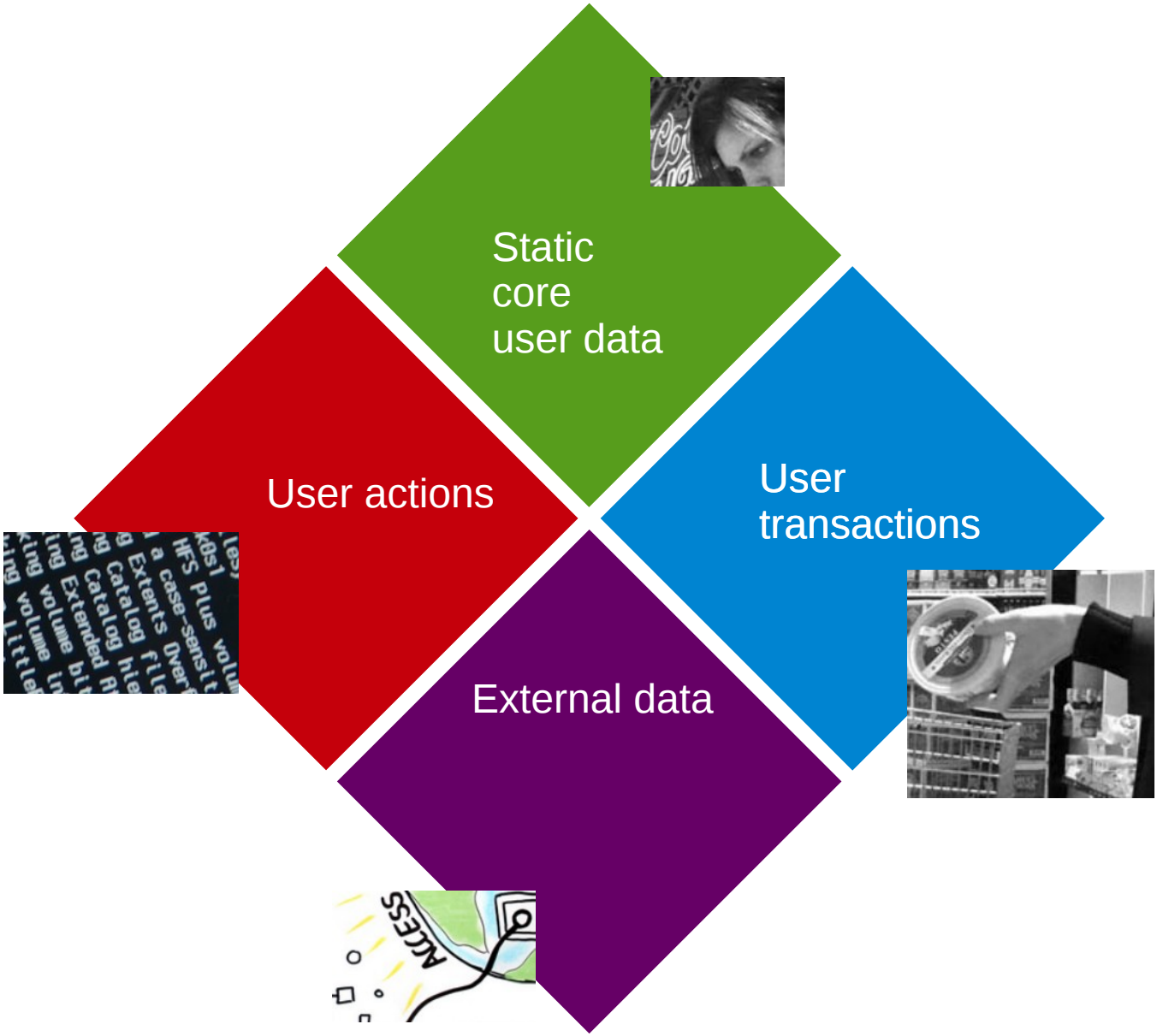

Static
core
user data


Static
core
user data
User
transactions

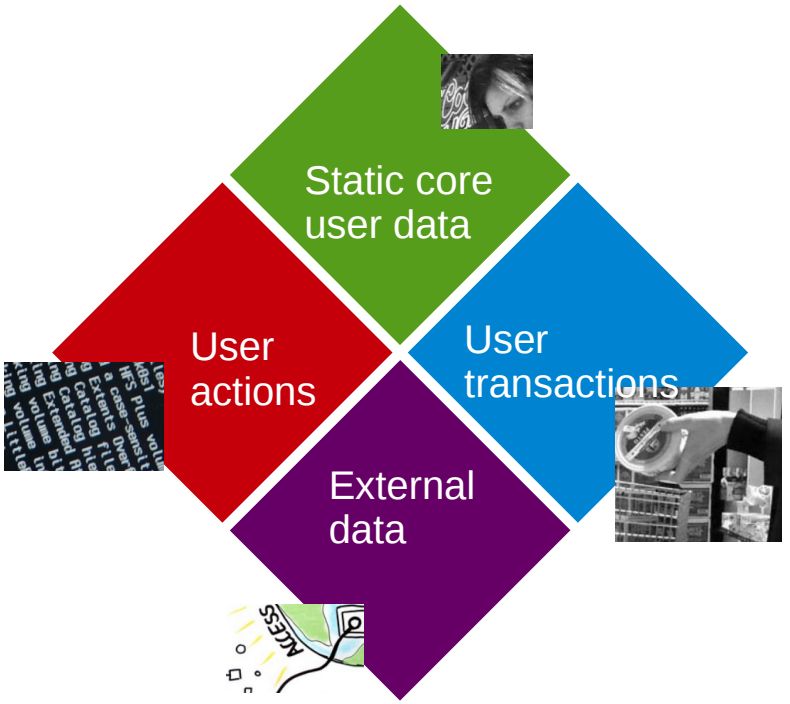

Image by Stoltze: http://www.flickr.com/photos/stoltze/2269137307/

By bareform, http://www.flickr.com/photos/bareform/2483573213/

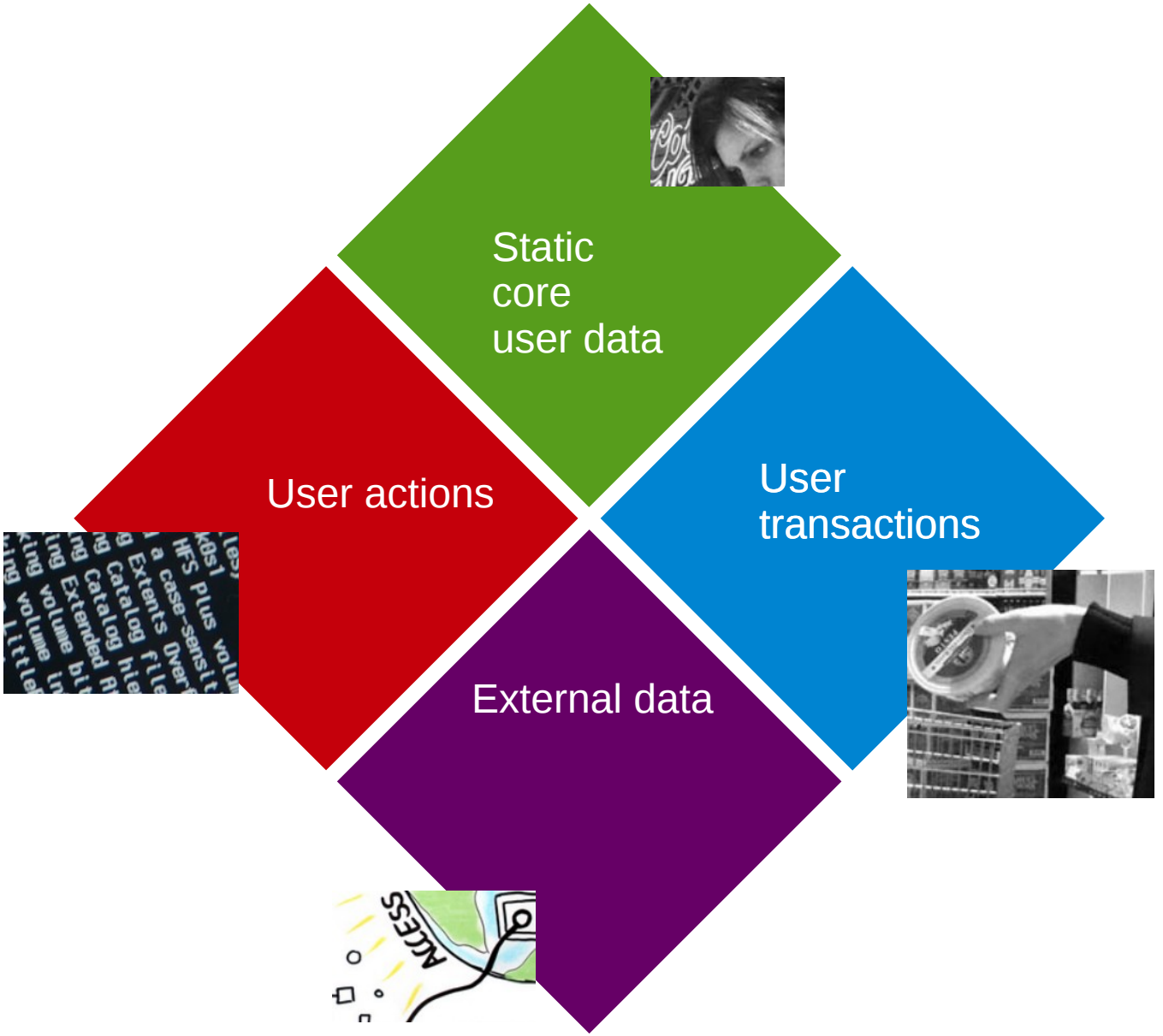

Static
core
user data
User actions
User
transactions


Static
core
user data
User actions
User
transactions
External data

Data volume



Data quality


Static
core
user data
User actions
User
transactions
External data

Discover content without effort by …

Showing content that I might like based on:

my personal interests and tags.

talks I already indicated as “want to attend.

abstracts and speaker info pages I read.

my prior search history and twitter messages.

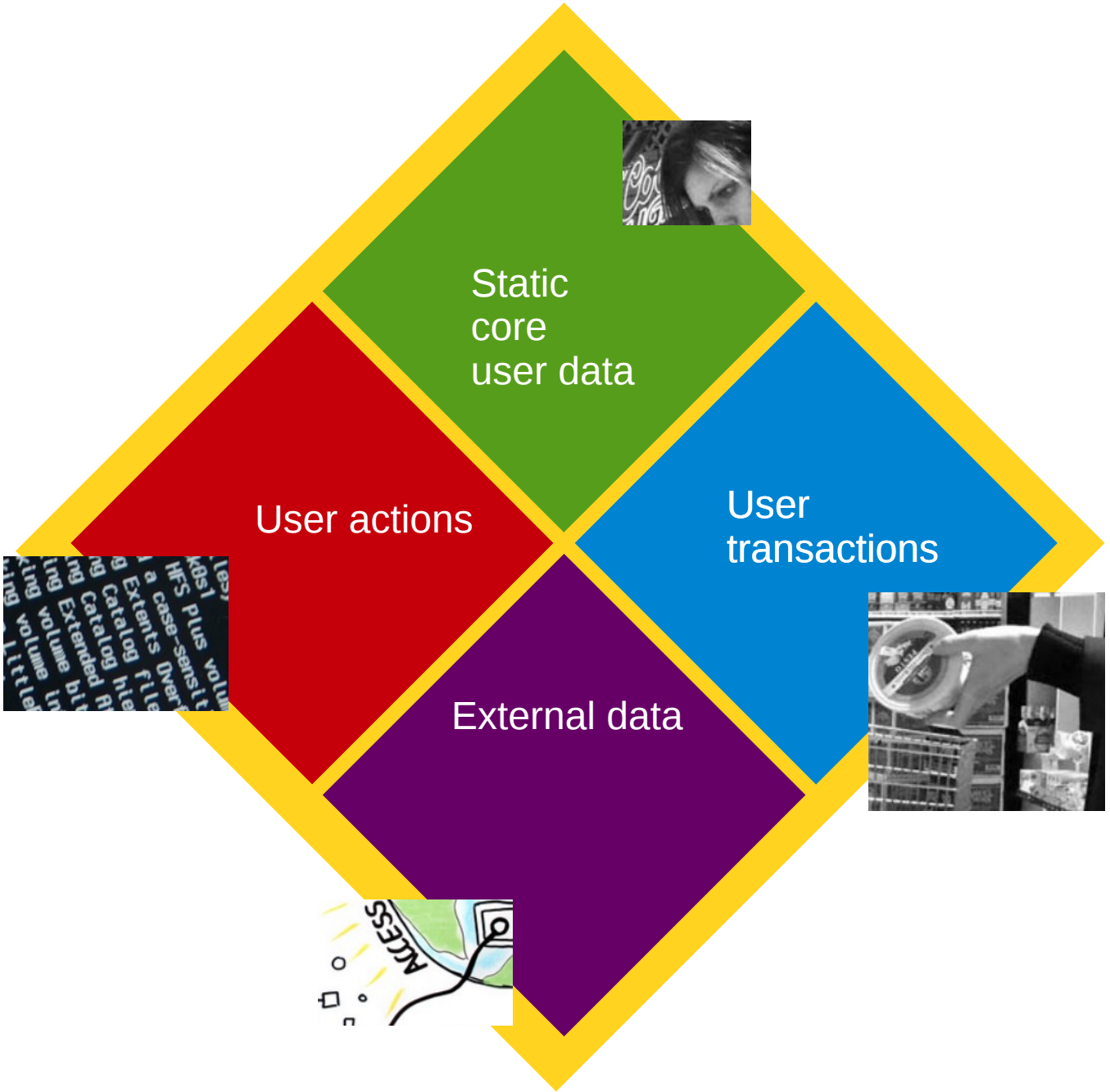

Static
core
user data
User actions
User
transactions
External data


Image by macfred64: http://www.flickr.com/photos/40275249@N04/4552190662/


Image by Garrett Rooney: http://www.flickr.com/photos/rooneg/276527837/

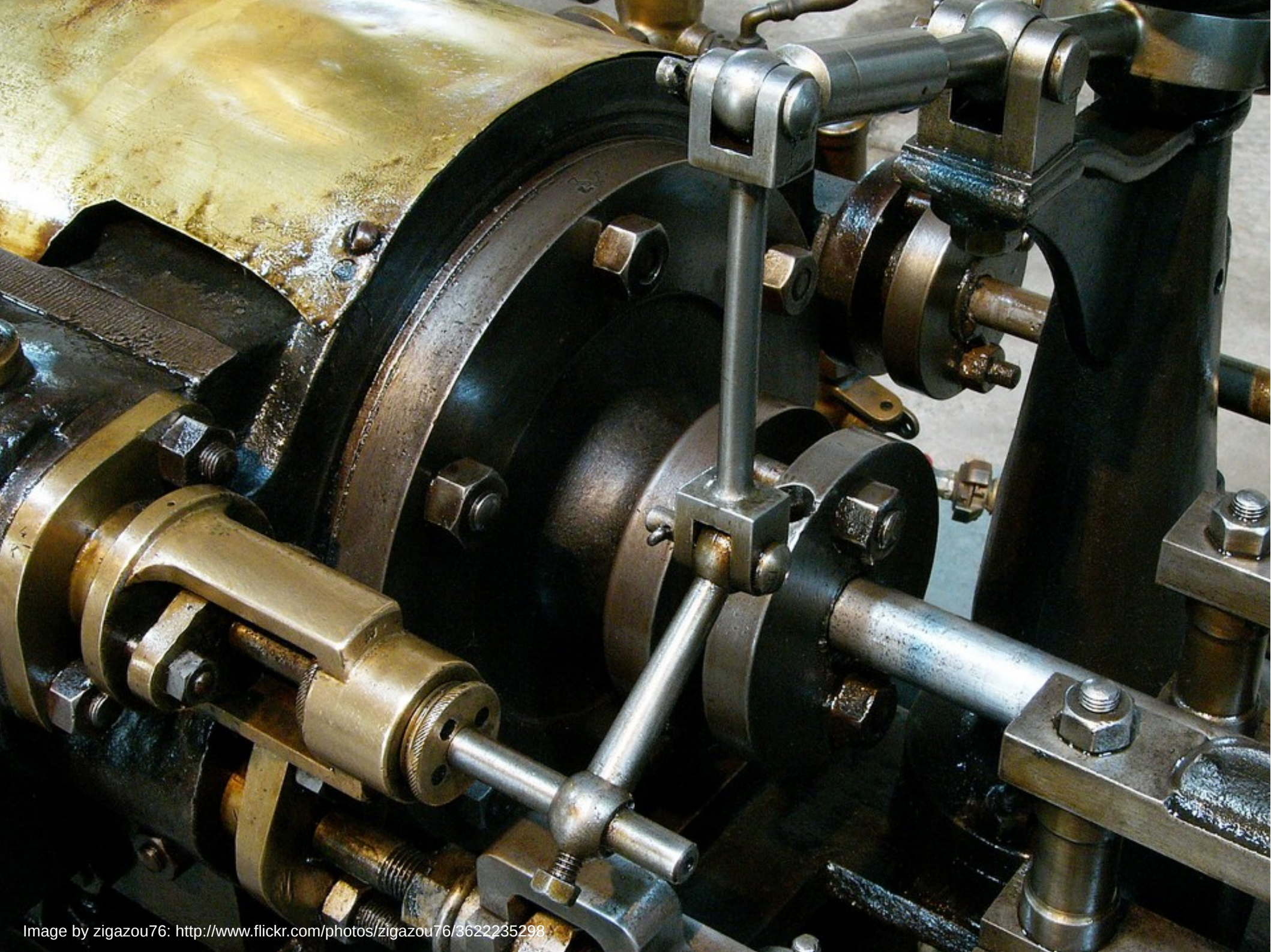


Image by zigazou76: http://www.flickr.com/photos/zigazou76/3622235298

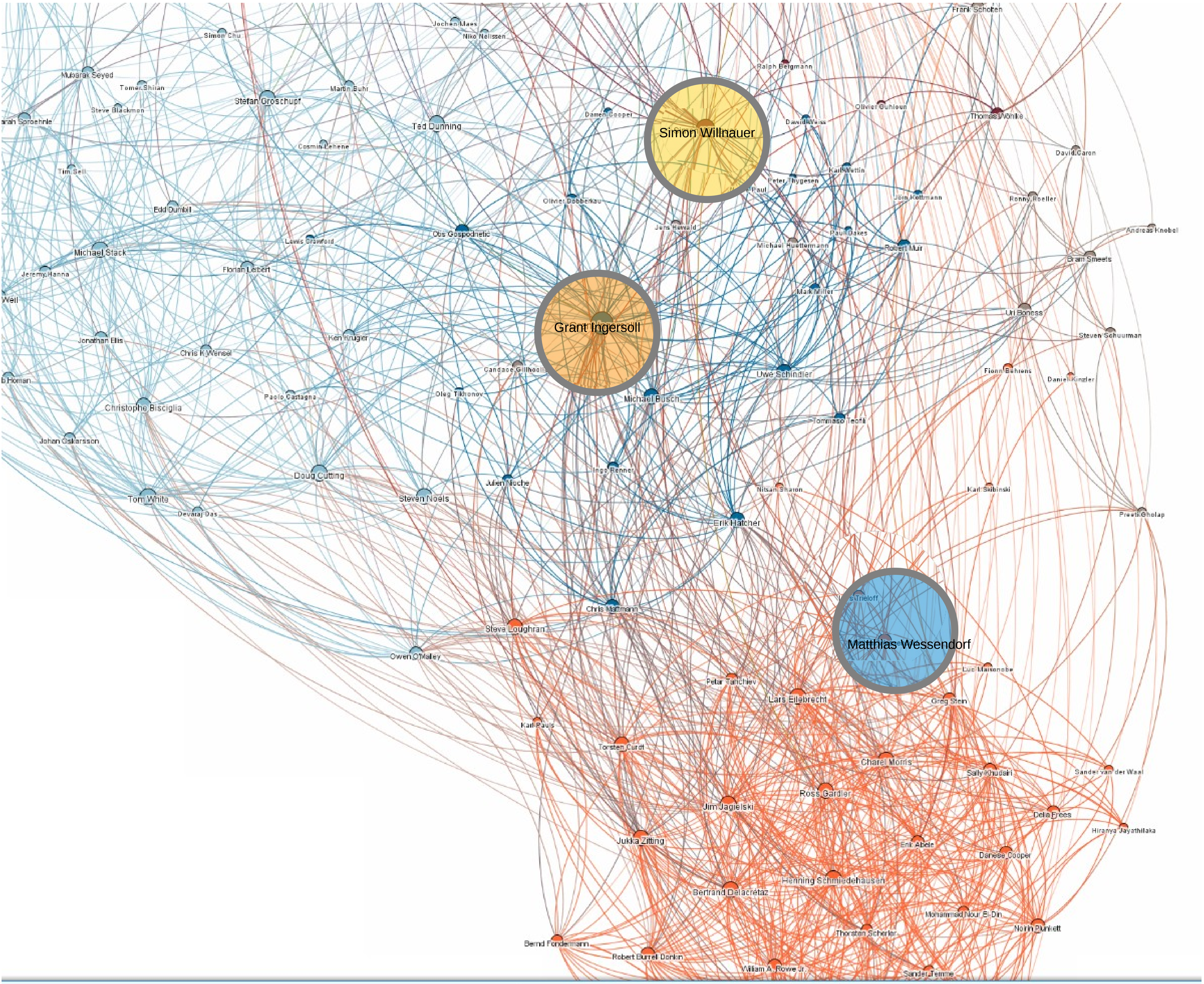

Simon Willnauer
Grant Ingersoll
Matthias Wessendorf
Frank Scholten
Jochen Maes
Niko Nelissen
Simon Chu
Ralph Beigmann
Mubarak Seyed
Tomer Shiran
Martin Buhr
Stefan Groschupf
Steve Blackmon
Olivier Guihoun
Thomas Vonite
Darren Cooper
David Weiss
Ted Dunning
Cosmin Lehene
David Caron
Tim Sell
Karl Wettin
Peter Thygesen
Jörn Kottmann
Ronny Roeller
Edd Dumbill
Olivier Dobberkau
Otis Gospodnetic
Jens Rewald
Paul Oakes
Andreas Knobel
Lewis Crawford
Michael Huettermann
Robert Muir
Michael Stack
Bram Smeets
Jeremy Hanna
Florian Leibert
Mark Miller
Weil
Uri Boness
Jonathan Ellis
Ken Krugler
Steven Schuurman
Chris K Wensel
Candace Gillhoolie
Uwe Schindler
Fionn Behrens
Daniel Kinzler
Paolo Castagna
Oleg Tikhonov
Michael Busch
Christophe Bisciglia
Tommaso Teofili
Johan Oskarsson
Ingo Renner
Doug Cutting
Julien Nioche
Nissan Sharon
Karl Suibinski
Tom White
Steven Noels
Devaraj Das
Erik Hatcher
Preeti Gholap
Chris Mattmann
Steve Loughran
Owen O'Malley
Luc Maisonobe
Petar Tahchiev
Lars Eilebrecht
Greg Stein
Karl Pauls
Torsten Curdt
Charel Morris
Sally Khudairi
Sander van der Waal
Ross Gardler
Jim Jagielski
Della Frees
Hiranya Jayathilaka
Jukka Zitting
Erik Abele
Danese Cooper
Henning Schmiedehausen
Bertrand Delacretaz
Mohammad Nour El-Din
Thorsten Scherler
Niclas Plunkett
Bernd Fondermann
Robert Burrell Donkin
William A. Rowe Jr.
Sander Temme


Image by macfred64:
http://www.flickr.com/photos/40275249@N04/4552190662/


Image by Garrett Rooney:
http://www.flickr.com/photos/rooneg/276527837/

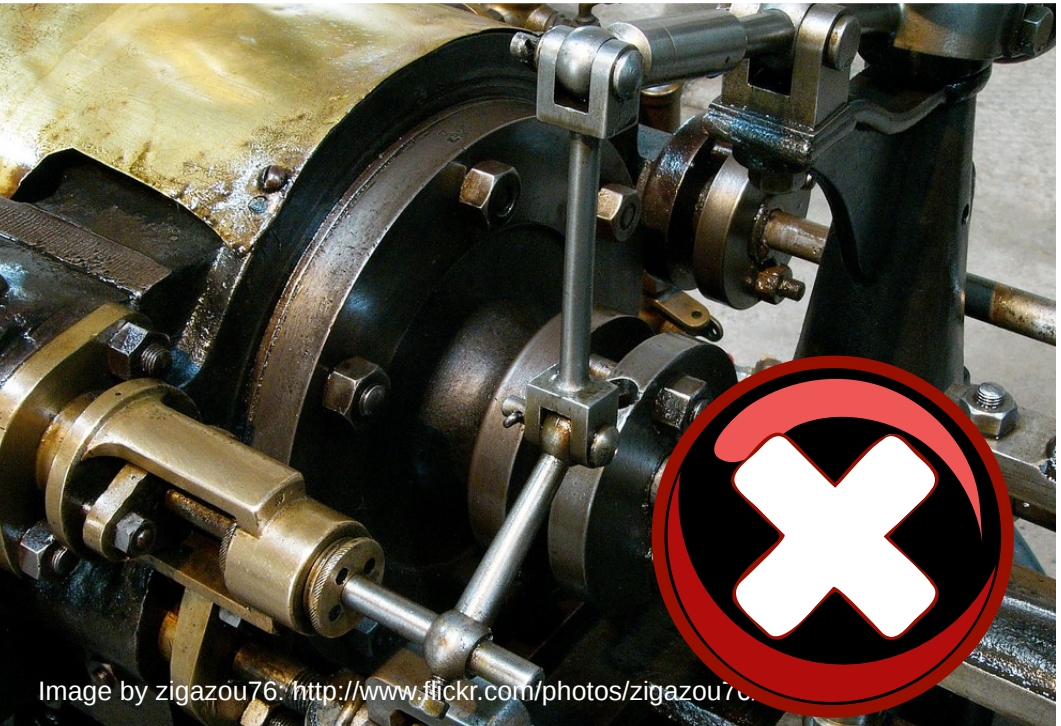
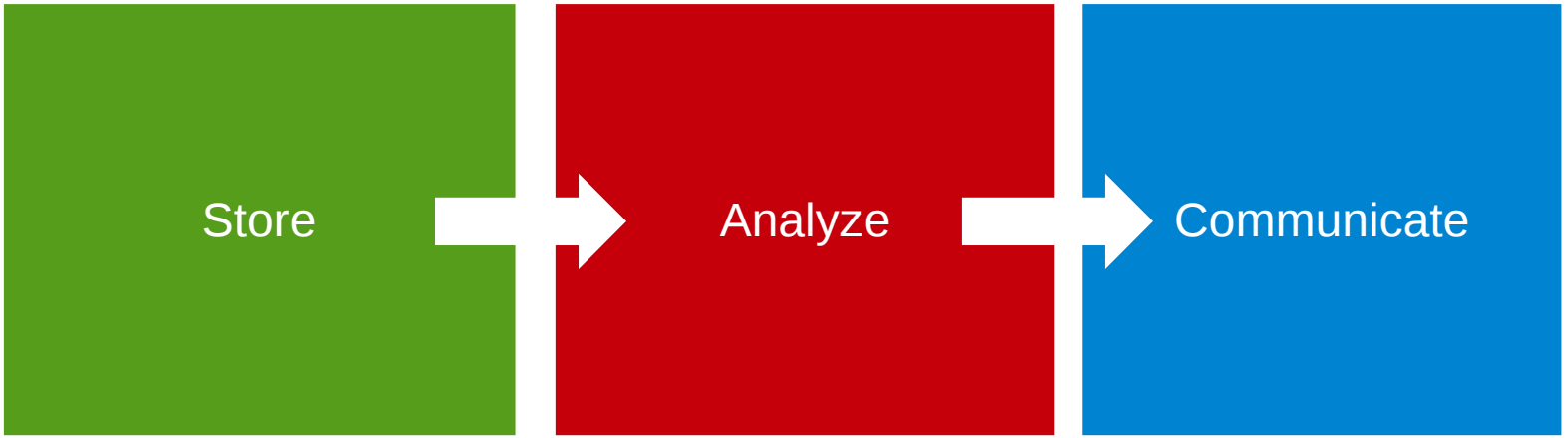


Image by zigazou76: http://www.flickr.com/photos/zigazou7

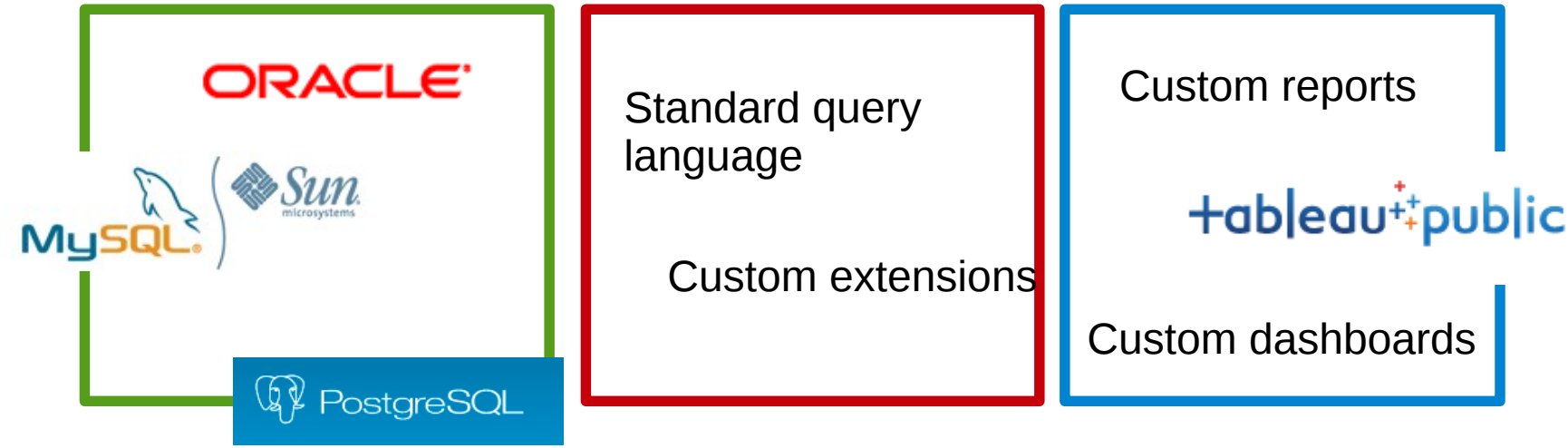

Isabel Drost
Simon Willnauer
Grant Ingersoll
Lars Trieloff
Matthias Wessendorf

Analysis tool-set


Store
Analyze
Communicate

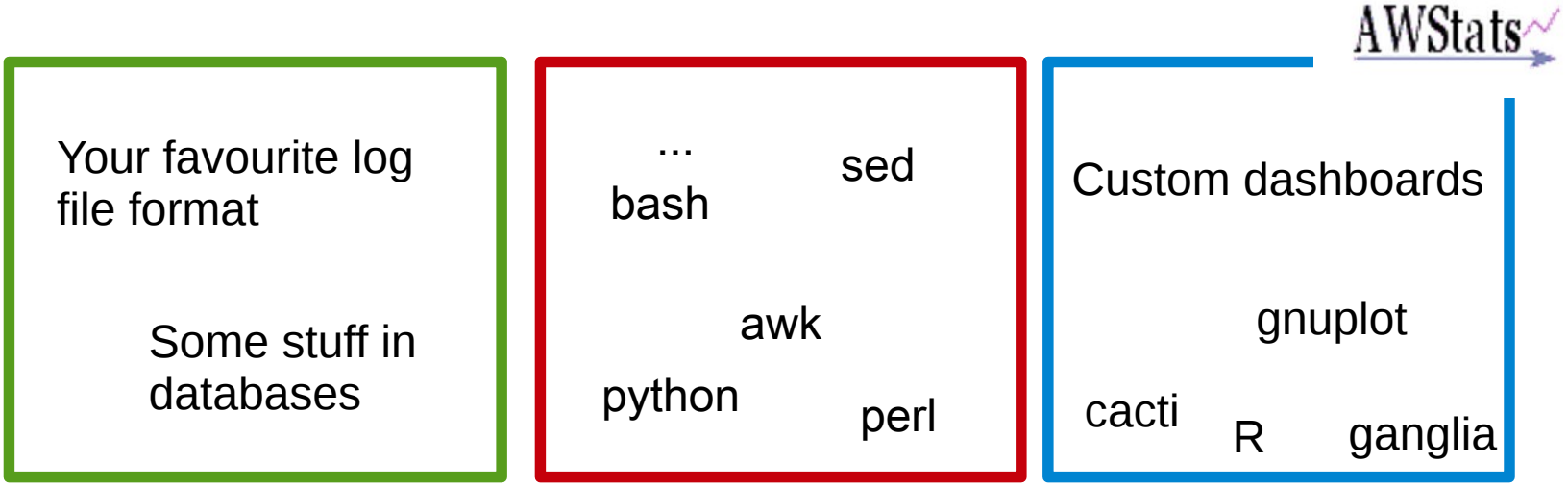
Structured relational data

- Attendee database.
- Speaker information.
- User transactions.



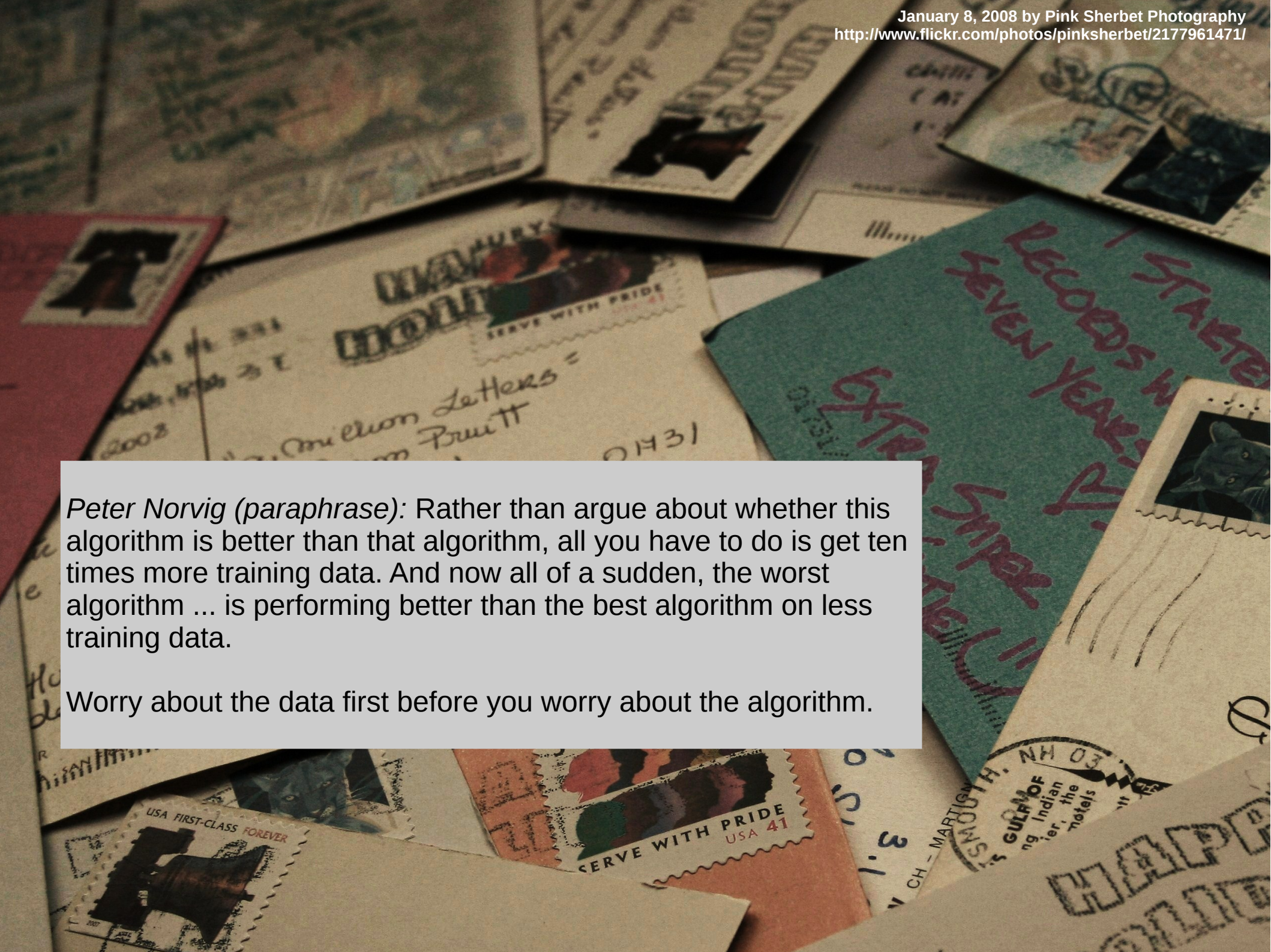
Extracting information from log files

- Access logs.
- Health check results.
- Response time logs.



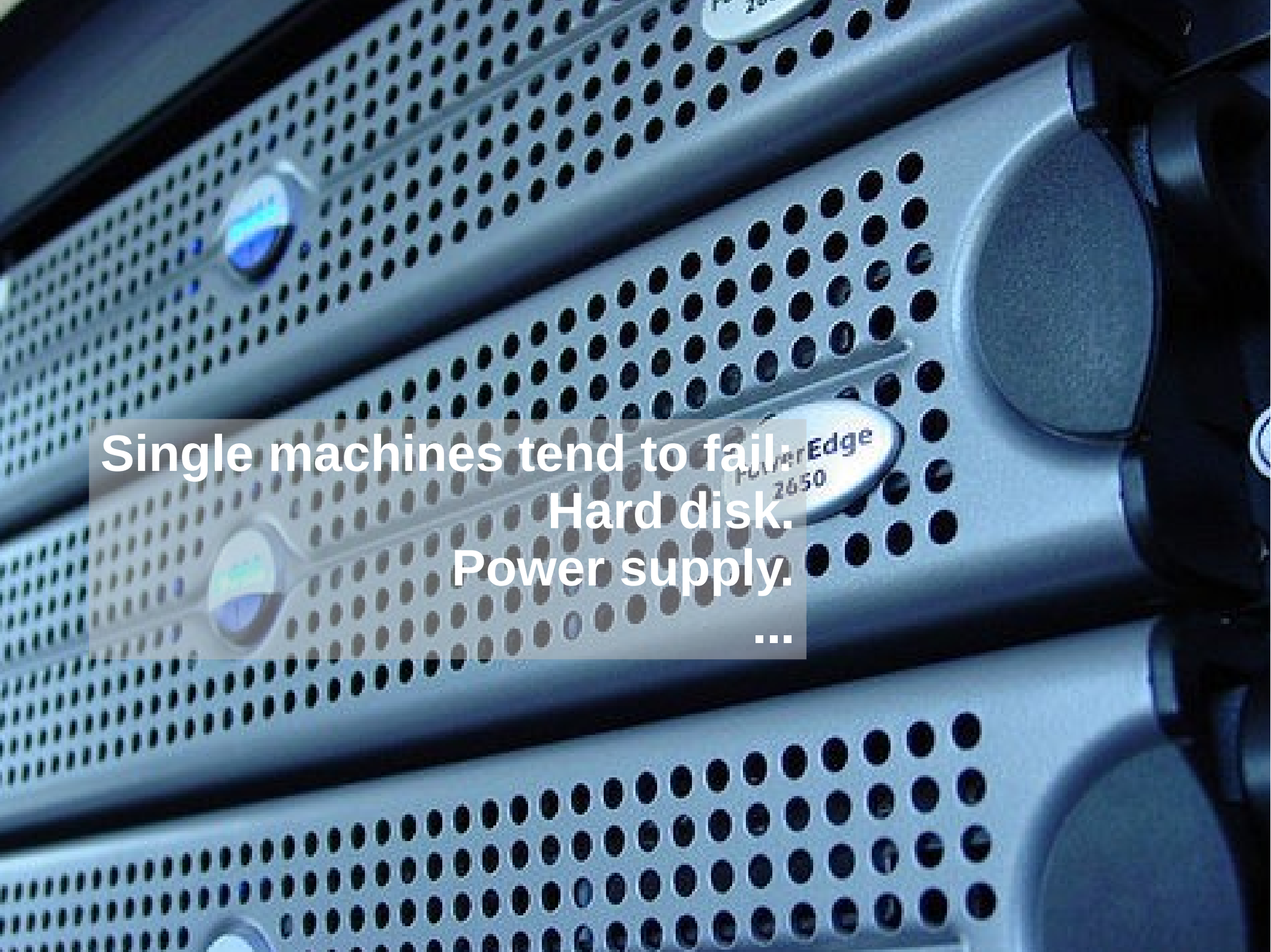
What about...

- Learning what users look for from search logs?
  - Do they query by room name?
  - … by speaker name?
  - … by tag?
- Learning what content to show from clicks?
  - Show different site per user?
  - Generate links between talks automatically?



Peter Norvig (paraphrase): Rather than argue about whether this algorithm is better than that algorithm, all you have to do is get ten times more training data. And now all of a sudden, the worst algorithm ... is performing better than the best algorithm on less training data.

Worry about the data first before you worry about the algorithm.



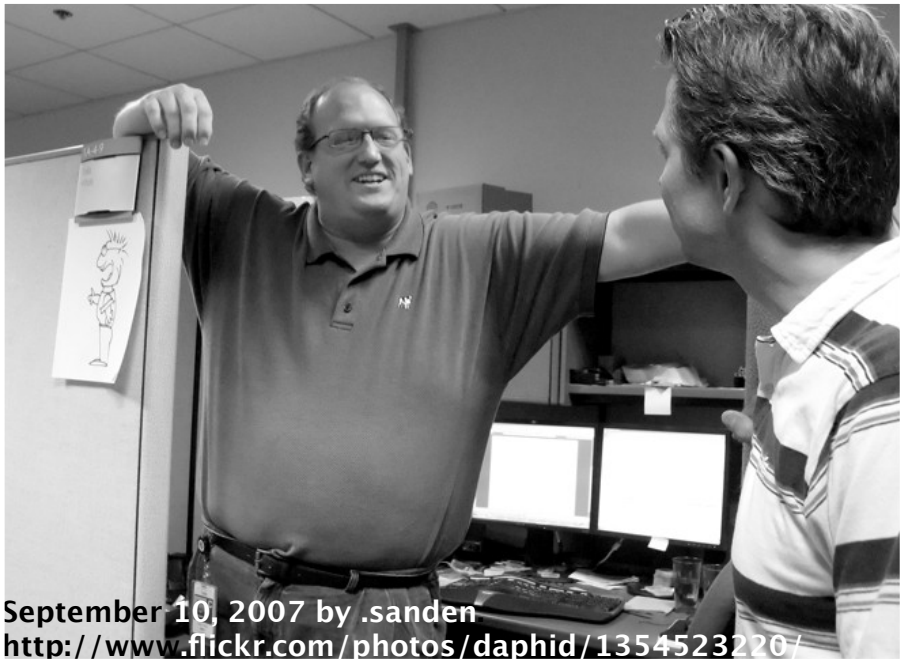
Single machines tend to fail.

Hard disk.

Power supply.

...

Typical developer



- Has never dealt with large (petabytes) amount of data.
- Has no thorough understanding of parallel programming.
- Has no time to make software production ready.

February 29, 2008 by Thomas Claveirole
http://www.flickr.com/photos/thomasclaveirole/2300932656/

http://www.flickr.com/photos/jaaronfarr/3384940437/
March 25, 2009 by jaaron

May 1, 2007 by danny angus
http://www.flickr.com/photos/killerbees/479864437/

http://www.flickr.com/photos
/jaaronfarr/3385756482/
March 25, 2009 by jaaron

http://www.flickr.com/photos/cspowers/282944734/ by cspowers on October 29, 2006

Easy distributed programming.

Well known in industry and research.

Scales well beyond 1000 nodes.

TM

Feb '03 first Map Reduce library @ Google

Oct '03 GFS Paper

Dec '04 Map Reduce paper

Dec '05 Doug reports that nutch uses map reduce

Feb '06 Hadoop moves out of nutch

Apr '07 Y! running Hadoop on 1000 node cluster

Jan '08 Hadoop made an Apache Top Level Project

Assumptions:

Data to process does not fit on one node.
Each node is commodity hardware.
Failure happens.

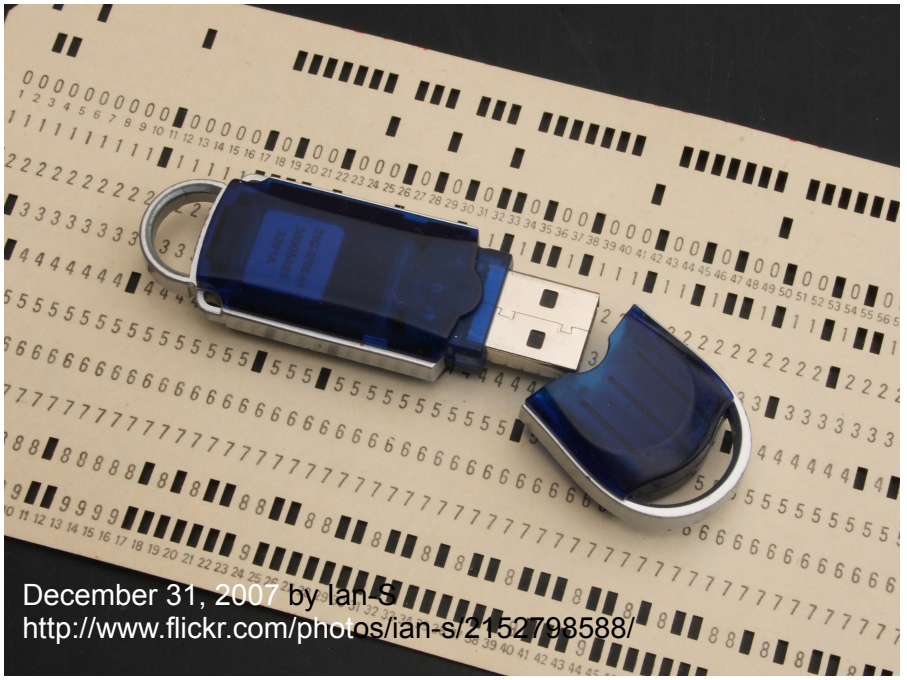
January 16, 2005 by JohnSeb
http://www.flickr.com/photos/johnseb/3425464/

Ideas:

Distribute filesystem.
Built in replication.
Automatic failover in case of failure.

Assumptions:

Distributed computation is easy.
Moving computation is cheap.
Moving data is expensive.



Ideas:

Move computation to data.
Write software that is easy to distribute.

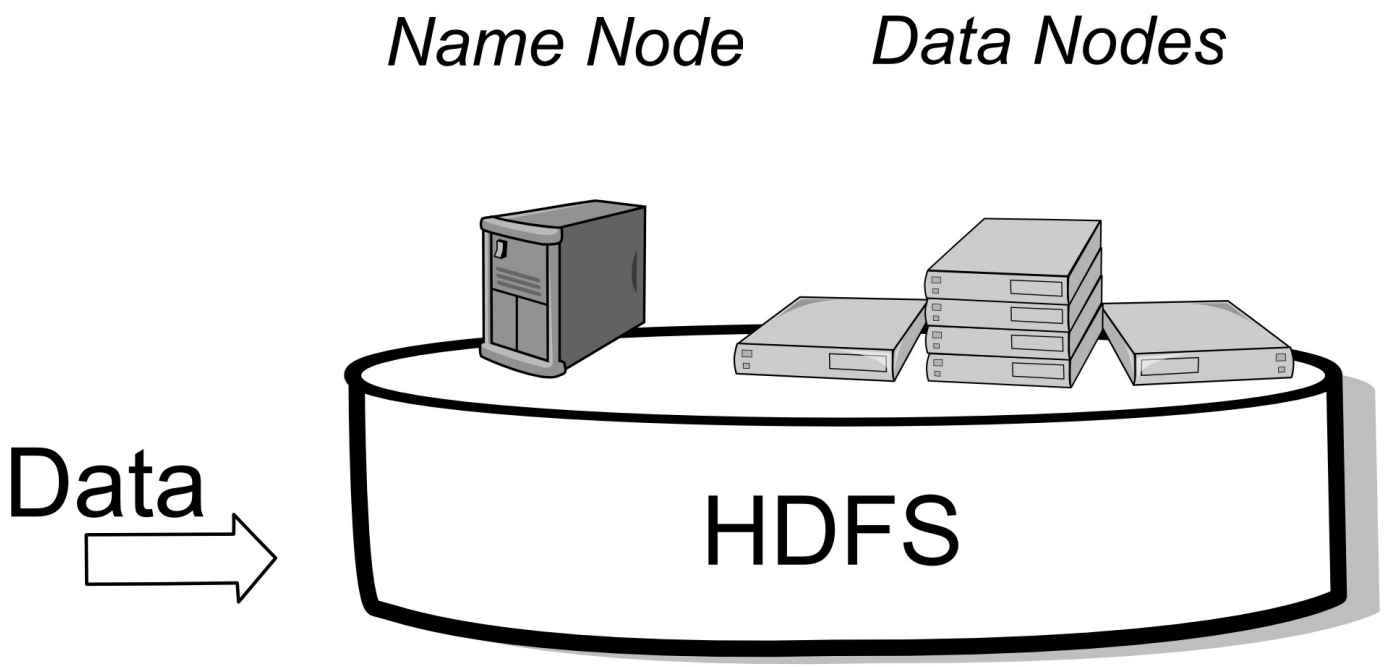
Assumptions:

Systems run on spinning hard disks.
Disk seek >> disk scan.

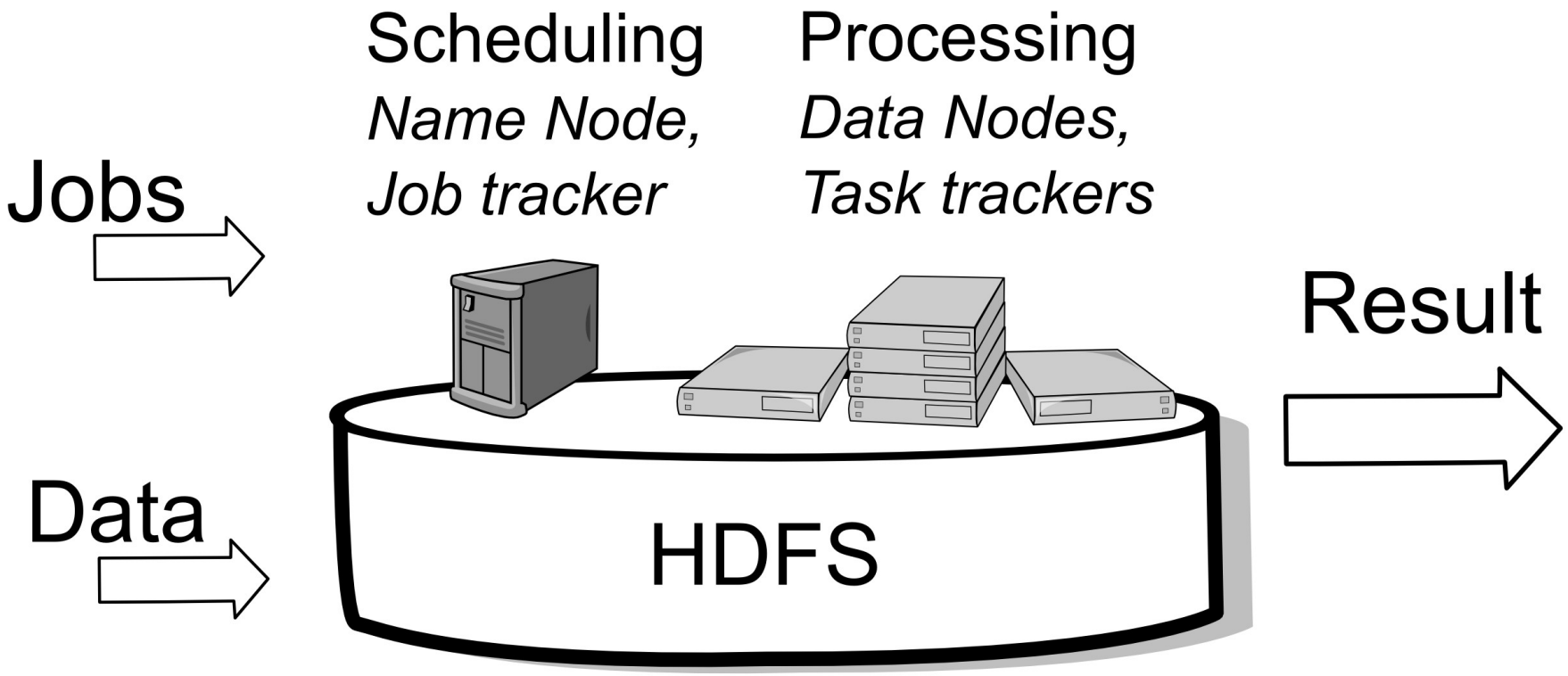
August 27, 2007 by Stuart Bryant
http://www.flickr.com/photos/stuartbryant/1249649001/

Ideas:

Improve support for large files.
File system API makes scanning easy.



(Graphics: Thanks to Thilo.)



(Graphics: Thanks to Thilo.)

```
<?xml version="1.0" encoding="UTF-8"?>
<opml version="1.0" >
 <head>
  <text></text>
 </head>
 <body>
  <outline htmlUrl="http://eventseer.net" title="EventSeer - A Digital Library of Call for Papers" useC
alDefault" version="RSS" type="rss" xmlUrl="http://eventseer.net/feeds/main/rss.xml" id="312053548" tex
tseer.net" />
  <outline isOpen="false" id="669809145" text="Silent" >
   <outline htmlUrl="http://www.theserverside.com" title="TheServerSide.com: Patterns" useCustomFetchIn
ersion="RSS" type="rss" xmlUrl="http://www.theserverside.com/rss/theserverside-j2eepatterns-rss2.xml" i
taining up-to-date news, discussions, patterns, resources, and media" />
   <outline htmlUrl="http://chadwa.wordpress.com" title="Chad's Search Blog" useCustomFetchInterval="fa
S" type="rss" xmlUrl="http://chadwa.wordpress.com/feed/" id="545368194" text="Chad's Search Blog" descr
" />
   <outline htmlUrl="http://www.find23.net/Site/Blog/Blog.html" title="My Blog" useCustomFetchInterval=
"RSS" type="rss" xmlUrl="http://www.find23.net/Site/Blog/rss.xml" id="1620106192" text="My Blog" descri
   <outline htmlUrl="http://emotion.inrialpes.fr/~dangauthier/blog" title="Yet Another Machine Learning
eMode="globalDefault" version="RSS" type="rss" xmlUrl="http://emotion.inrialpes.fr/~dangauthier/blog/fe
g" />
   <outline htmlUrl="http://ml.typepad.com/machine_learning_thoughts/" title="Machine Learning Thoughts
="globalDefault" version="RSS" type="rss" xmlUrl="http://ml.typepad.com/machine_learning_thoughts/rss.x
etical and practical aspects of Machine Learning." />
   <outline htmlUrl="http://yaroslavvb.blogspot.com/" title="Machine Learning, etc" useCustomFetchInter
ion="RSS" type="rss" xmlUrl="http://yaroslavvb.blogspot.com/feeds/posts/default" id="805998569" text="M
   <outline htmlUrl="http://ptufts.blogspot.com/" title="Pinhead's Progress" useCustomFetchInterval="fa
S" type="rss" xmlUrl="http://ptufts.blogspot.com/feeds/posts/default" id="1019393988" text="Pinhead's P
   <outline htmlUrl="http://resnotebook.blogspot.com/" title="Misc Research Stuff" useCustomFetchInterv
on="RSS" type="rss" xmlUrl="http://resnotebook.blogspot.com/feeds/posts/default" id="216193226" text="M
   <outline htmlUrl="http://absolutely-regular.blogspot.com/" title="Absolutely Regular" useCustomFetch
 version="RSS" type="rss" xmlUrl="http://absolutely-regular.blogspot.com/feeds/posts/default" id="17850
   <outline htmlUrl="http://atomai.blogspot.com/" title="Data Mining, Analytics and Artificial Intelli
Mode="globalDefault" version="RSS" type="rss" xmlUrl="http://atomai.blogspot.com/feeds/posts/default" i
nt in data mining, artificial intelligence, analytics, intelligent agents, semiconductors, distributing
siness Objects, Oracle, Intel, AMD, or Pentaho. Heuristic, Six Sigma, or CMM. Contractor or in-house. H
ail.com" />
```

```
isabel@h1349259:~$ more data/feeds.opml | grep -o "http://[0-9A-Za-z\-_\.]*" | s
ort | uniq --count | sort | tail
      3 http://agbs.kyb.tuebingen.mpg.de
      3 http://irgupf.com
      3 http://jeffsutherland.com
      4 http://ml.typepad.com
      4 http://weblogs.java.net
      4 http://www.gridvm.org
      4 http://yaroslavvb.blogspot.com
      5 http://feeds.feedburner.com
      6 http://blogsearch.google.com
     10 http://arxiv.org
```

```
pattern=”http://[0-9A-Za-z\-_\.]*”

grep -o "$pattern" feeds.opml | sort | uniq --count
```

TM

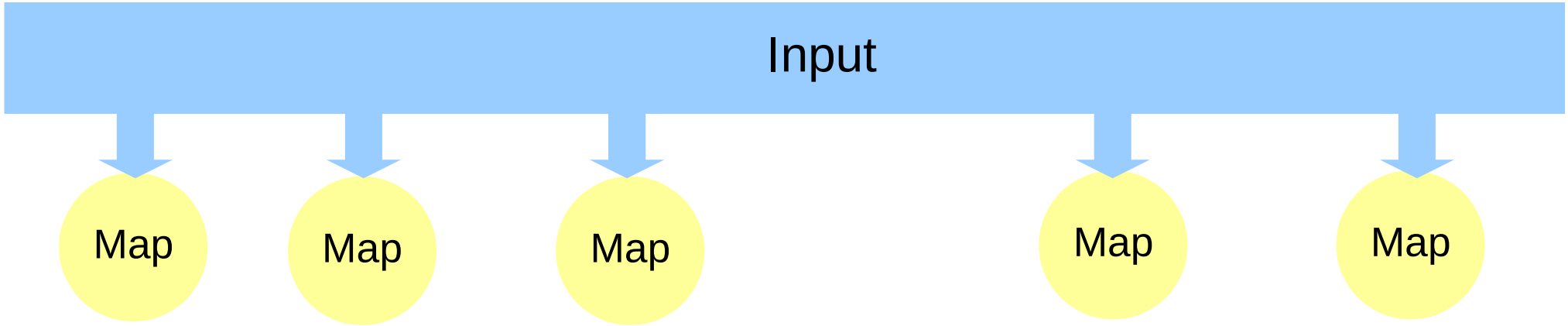
pattern=”http://[0-9A-Za-z\-_\.]*”

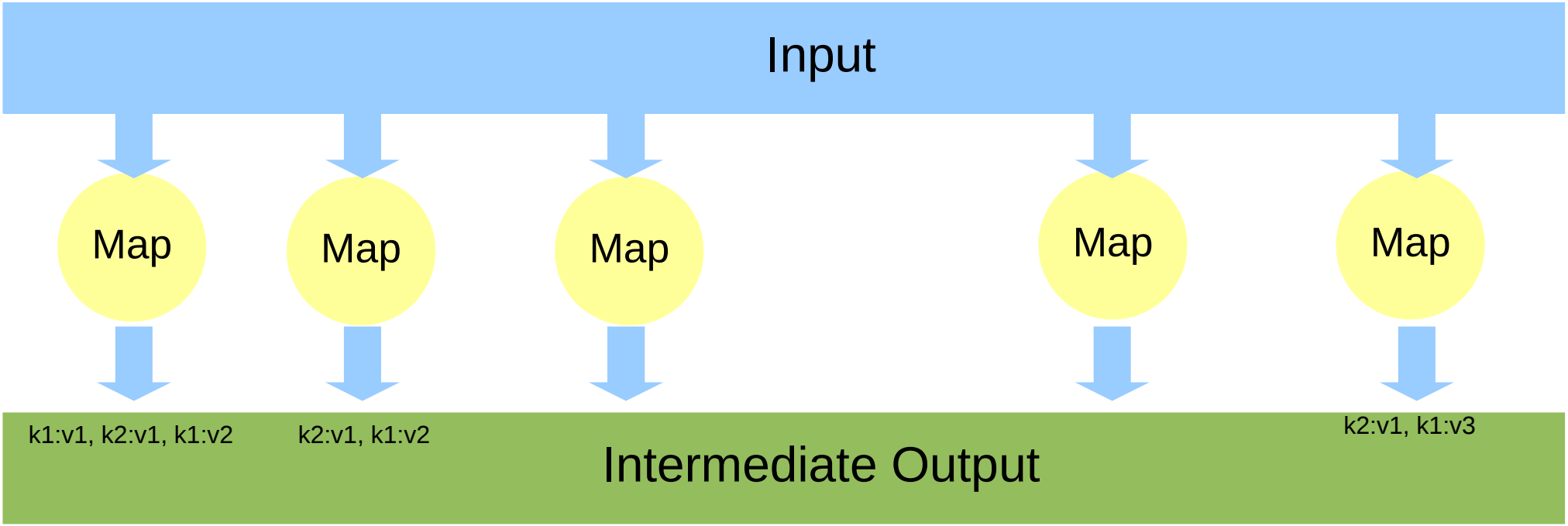
```
grep -o "$pattern" feeds.opml

M  A  P
```

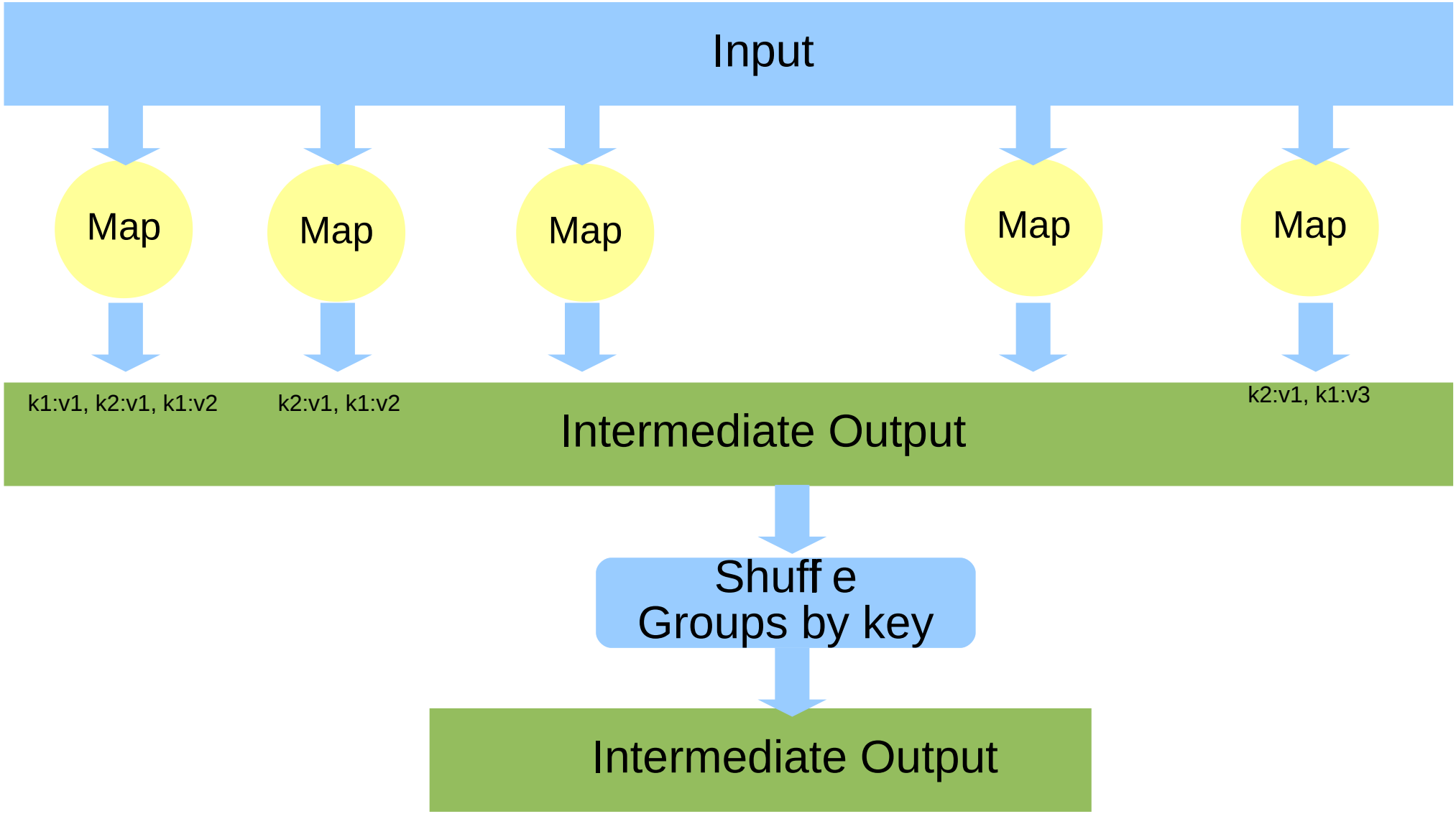
```
| sort

| SHUFFLE
```

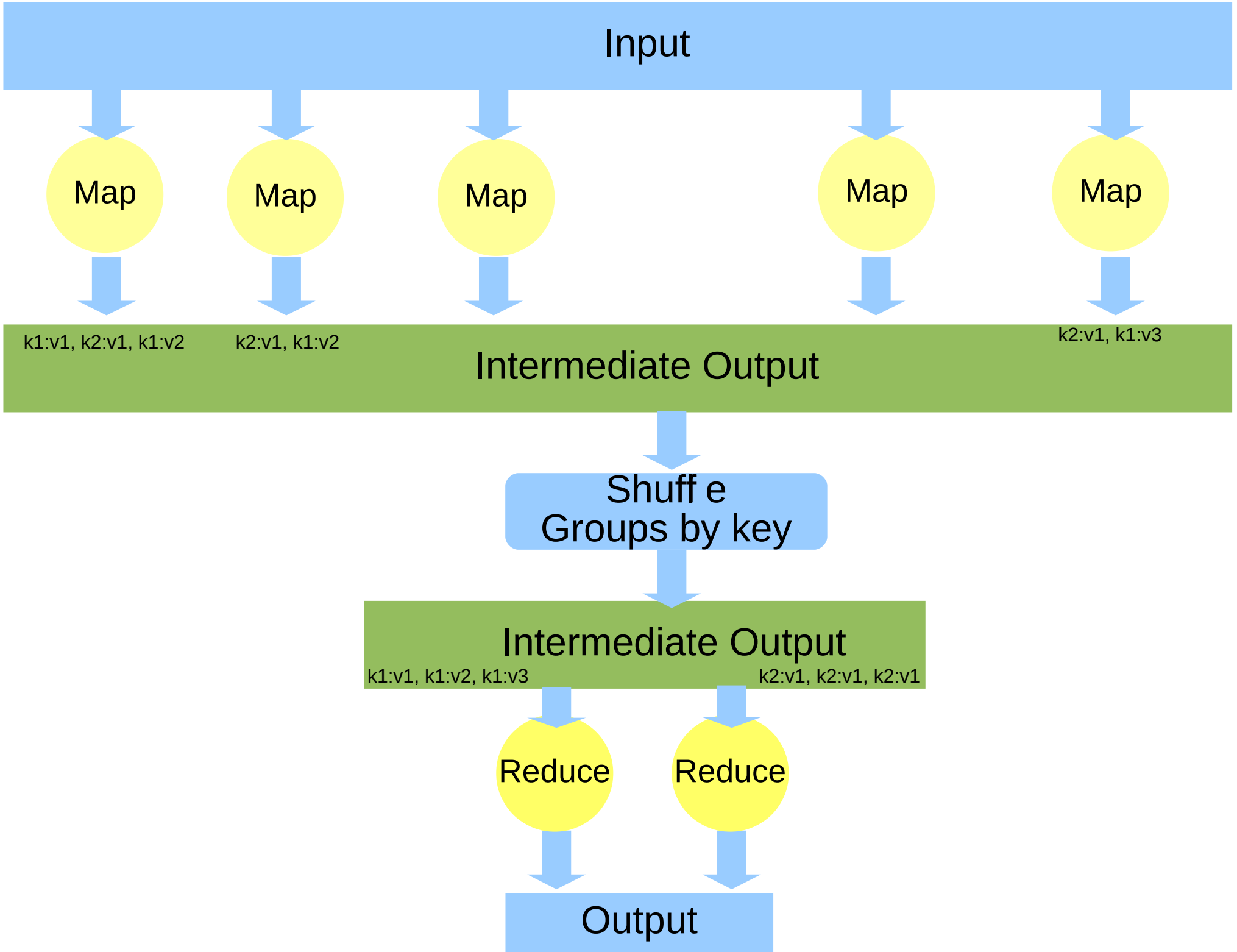
```
| uniq --count

| R E D U C E
```

TM


Input
Map
Map
Map
Map
Map

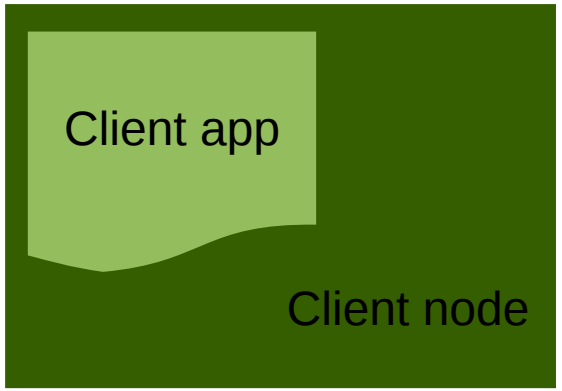

Input
Map
Map
Map
Map
Map
k1:v1, k2:v1, k1:v2
k2:v1, k1:v2
k2:v1, k1:v3
Intermediate Output


Input
Map
Map
Map
Map
Map
k1:v1, k2:v1, k1:v2
k2:v1, k1:v2
k2:v1, k1:v3
Intermediate Output
Shuff e
Groups by key
Intermediate Output


Input
Map
Map
Map
Map
Map
k1:v1, k2:v1, k1:v2
k2:v1, k1:v2
k2:v1, k1:v3
Intermediate Output
Shuff e
Groups by key
Intermediate Output
k1:v1, k1:v2, k1:v3
k2:v1, k2:v1, k2:v1
Reduce
Reduce
Output

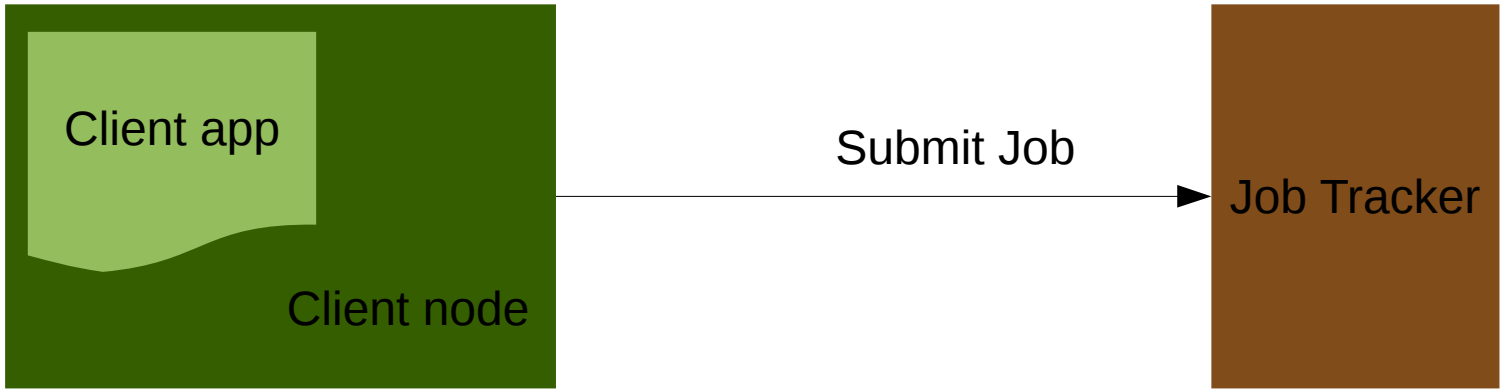
```
private IntWritable one = new IntWritable(1);
private Text hostname = new Text();


public void map(K key, V value, Context context) {
  String line = value.toString();
  StringTokenizer tokenizer = new StringTokenizer(line);
  while (tokenizer.hasMoreTokens()) {
    hostname.set(getHostname(tokenizer.nextToken()));
    context.write(hostname, one);
  }
}


public void reduce(K2 key, Iterable<V2> values,
  OutputCollector<K3, V2> output) {
  int sum = 0;
  while (values.hasNext()) {
    sum += values.next().get();
  }
  output.collect(key, new IntWritable(sum));
}
```

Anatomy of a map/reduce job



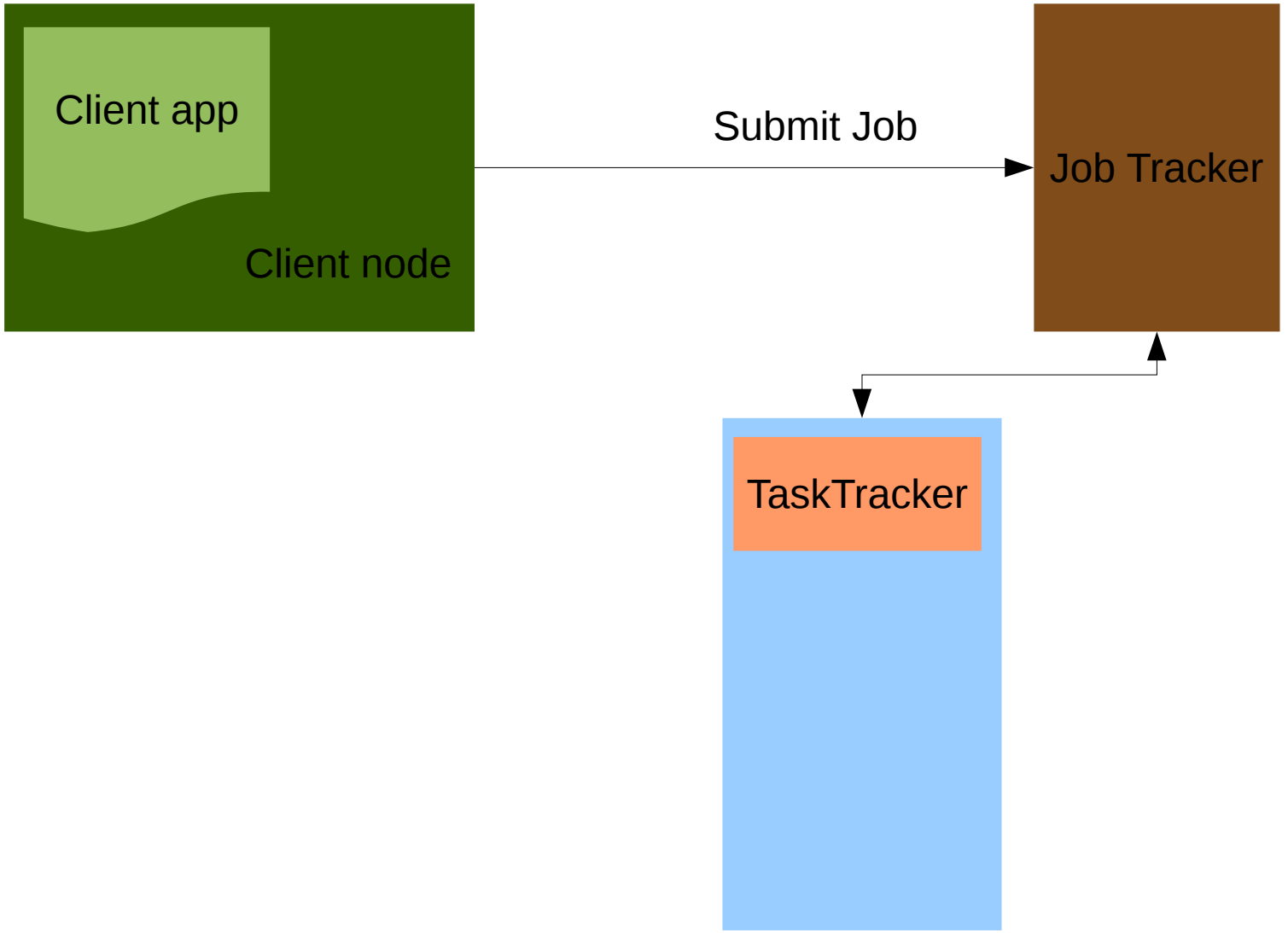
Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly

Anatomy of a map/reduce job



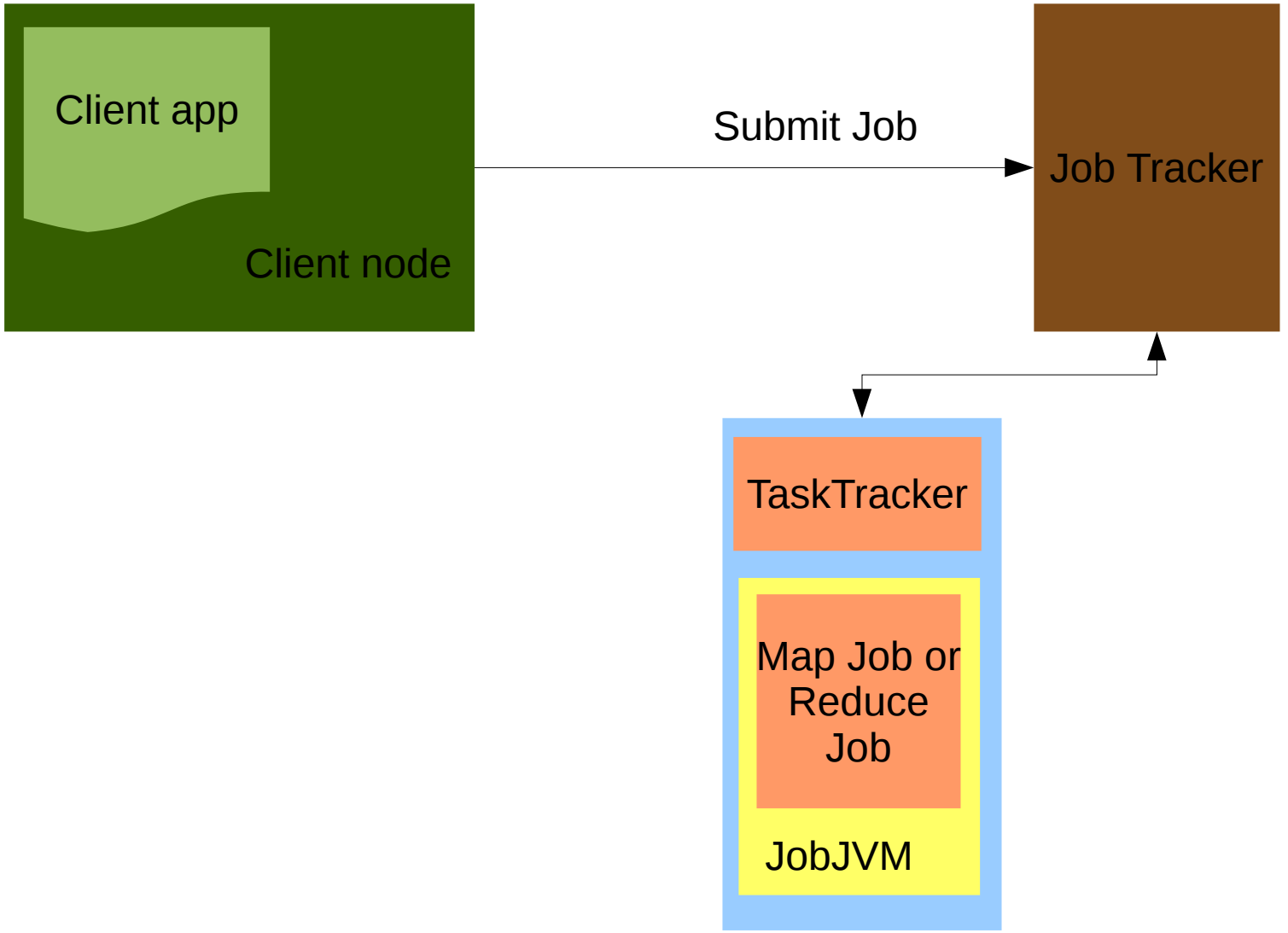
Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly

Anatomy of a map/reduce job



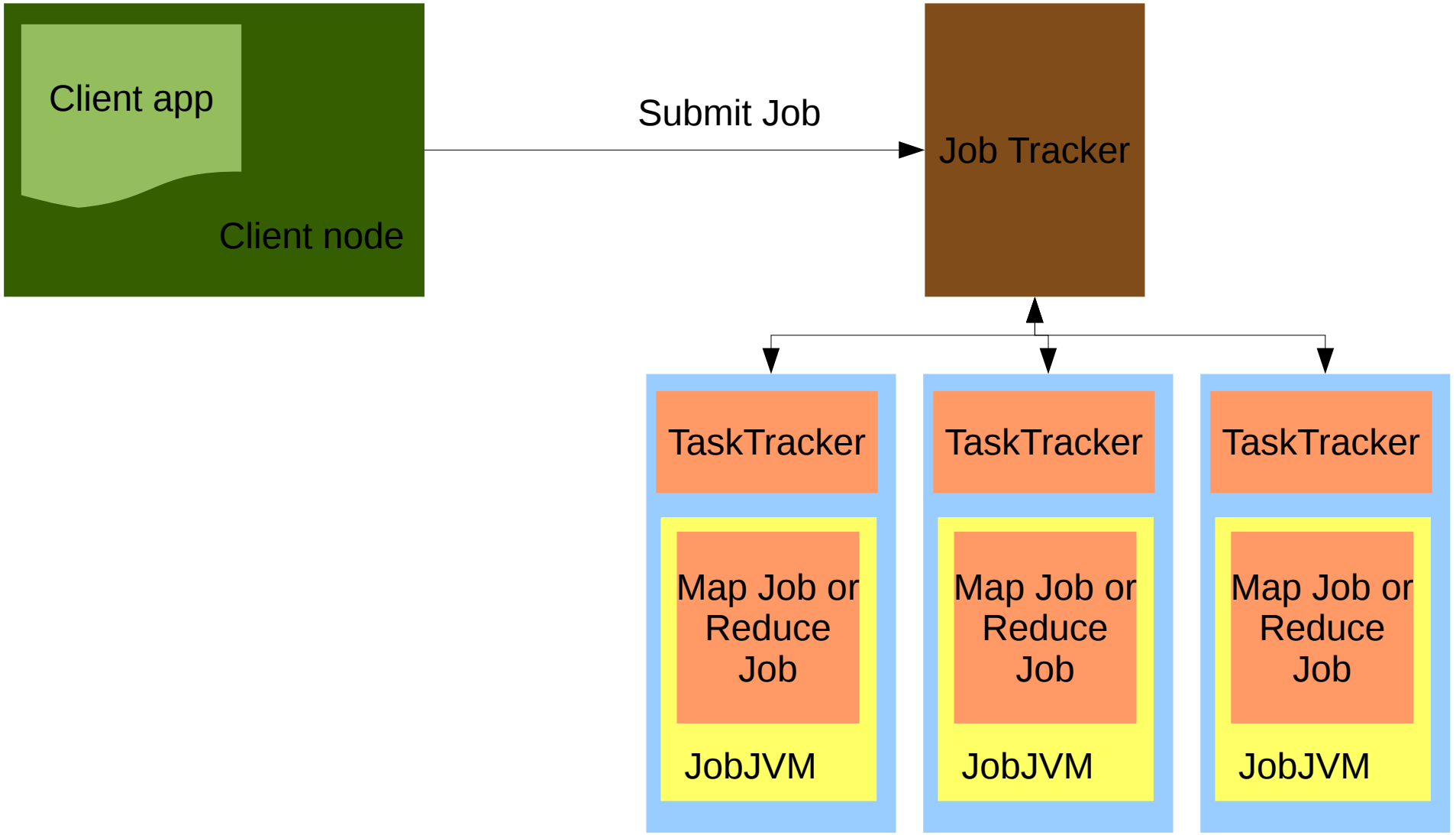
Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly

Anatomy of a map/reduce job



Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly

Anatomy of a map/reduce job



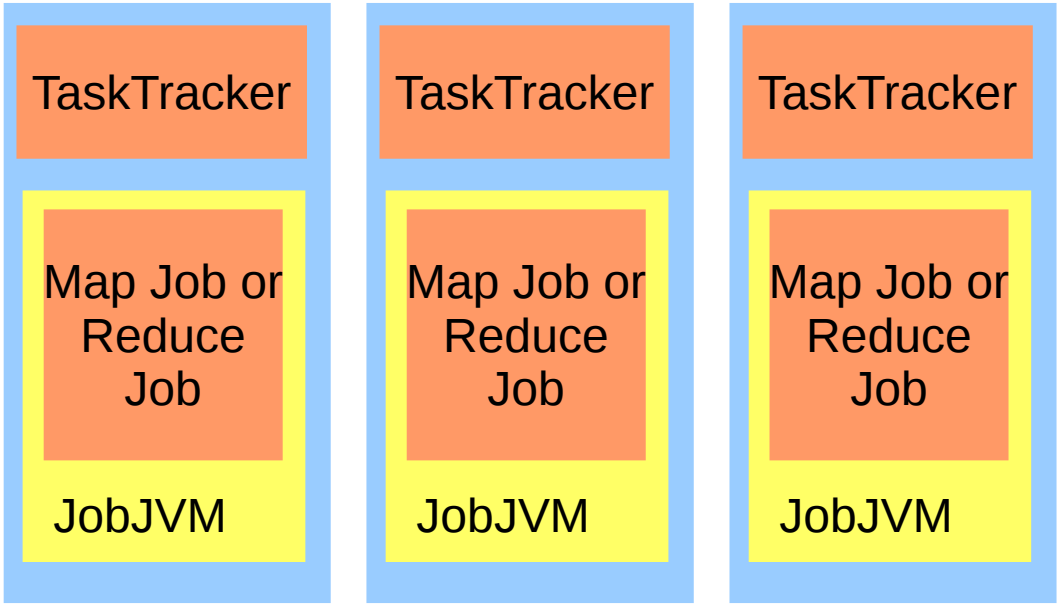
Slide inspired by: “Hadoop – The definitive guide”, Tom White, O'Reilly

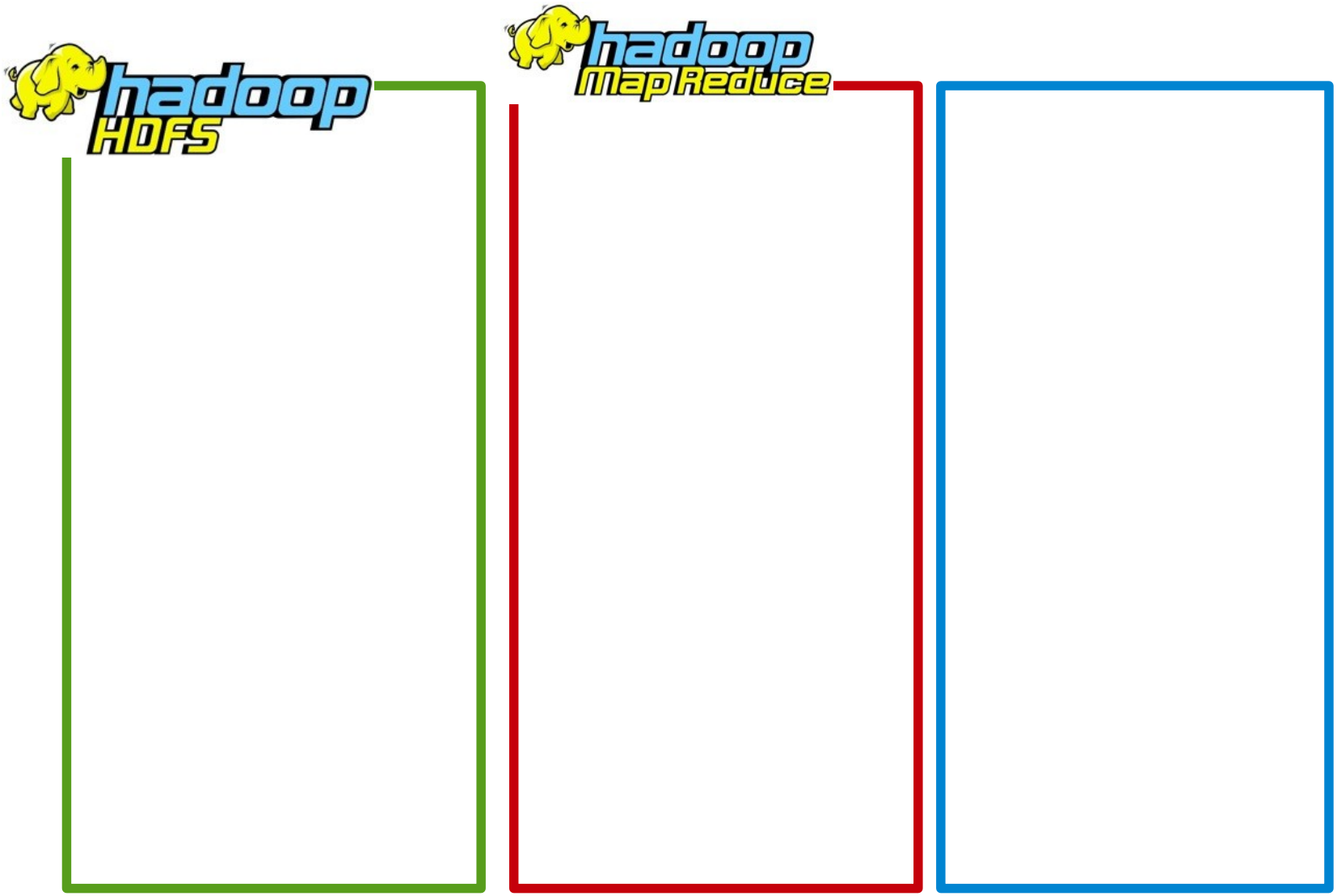
Properties of good map reduce citizens:

do not crash on increasing amount of input data.

do not overload external resources.

self contained.




hadoop
HDFS
hadoop
MapReduce

About MRUnit

MRUnit is a Java library that helps devel

Welcome to Apache HCatalog!

Hadoop Eclipse Plug-in

PigPen

PigPen is an eclipse plugin that helps users cı

About MRUnit

MRUnit is a Java library that helps devel

About MRUnit

MRUnit is a Java library that helps devel

About MRUnit

MRUnit is a Java library that helps devel


TM
H·BASE

Adds to Hadoop:

Random access to your data

Near-Realtime read/write

Host very large tables (billions of rows, millions of columns)

Google code protobuf
Protocol Buffers - Google's data interchange format

TM

TM

TM

Java API
Pipes C++ API
Streaming
Dumbo for Python
...

TM

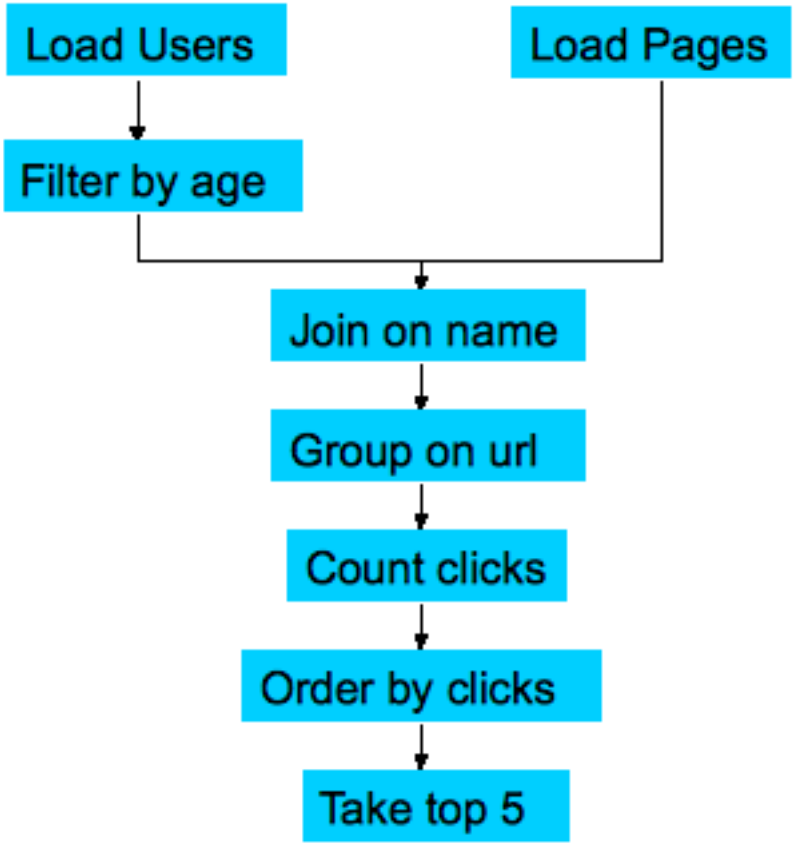
About MRUnit

MRUnit is a Java library that helps devel

TM

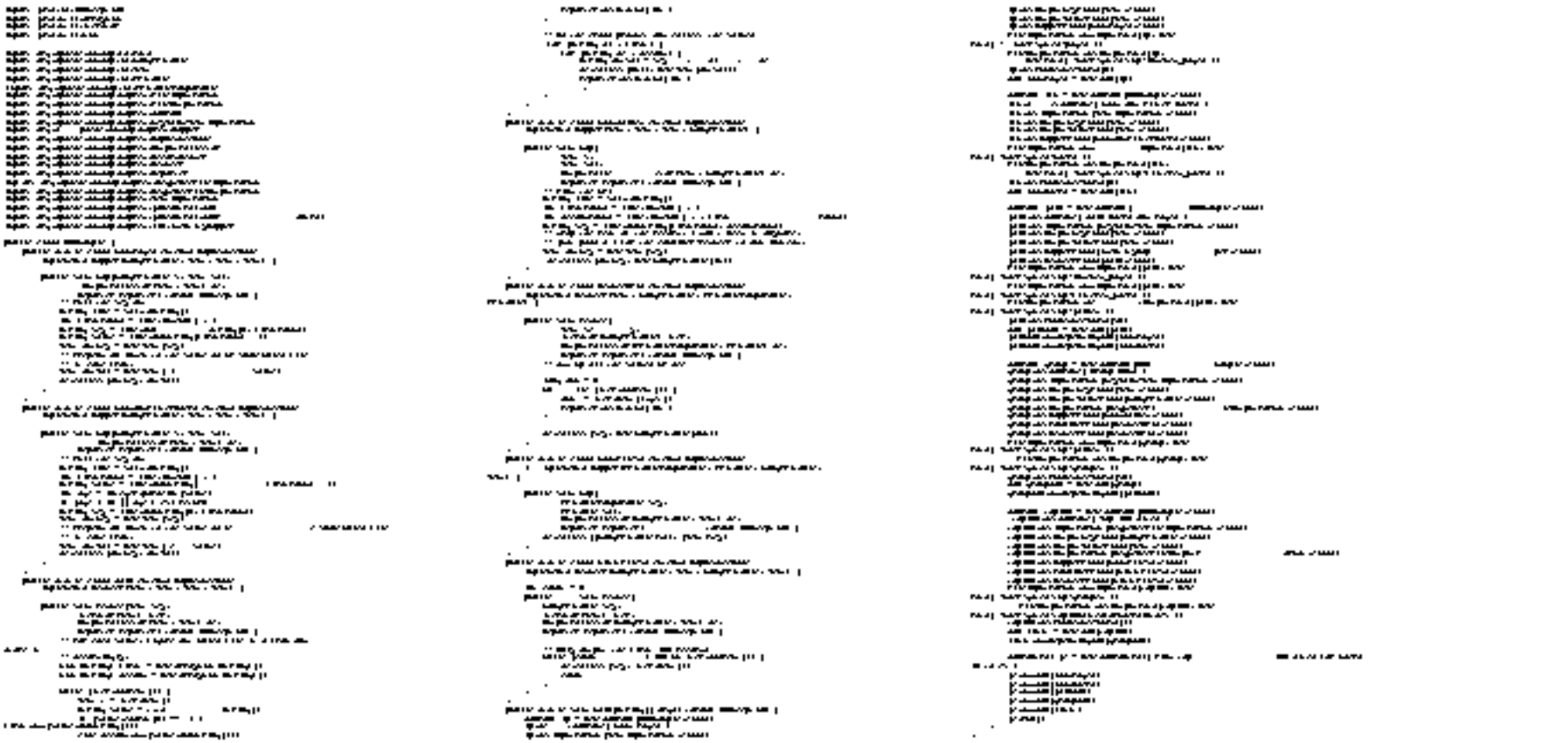
TM

Suppose you have user data in one file, website data in another, and you need to find the top 5 most visited pages by users aged 18 - 25.



Example from PIG presentation at Apache Con EU 2009

TM



Example from PIG presentation at Apache Con EU 2009

TM

```
Users = load 'users' as (name, age);
Fltrd = filter Users by
        age >= 18 and age <= 25;
Pages = load 'pages' as (user, url);
Jnd = join Fltrd by name, Pages by user;
Grpd = group Jnd by url;
Smmd = foreach Grpd generate group,
       COUNT(Jnd) as clicks;
Srtd = order Smmd by clicks desc;
Top5 = limit Srtd 5;
store Top5 into 'top5sites';
```

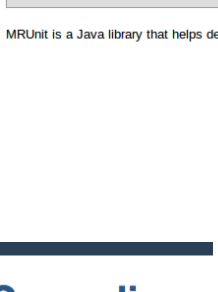
Example from PIG presentation at Apache Con EU 2009

TM

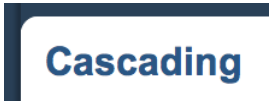
TM

TM

Java API
Pipes C++ API
Streaming
Dumbo for Python
...



TM



About MRUnit

MRUnit is a Java library that helps devel

TM

Java API
Pipes C++ API
Streaming
Dumbo for Python
...

TM

TM

About MRUnit

MRUnit is a Java library that helps devel

TM

TM

Cascading

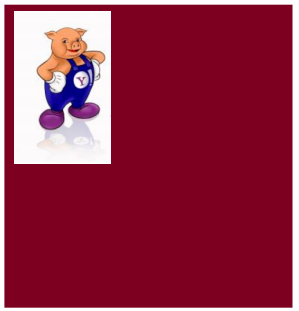
TM

TM

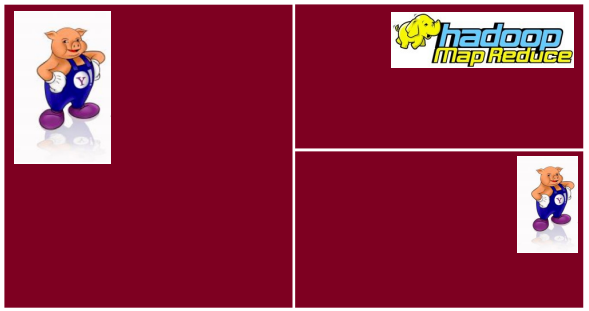

OOZIE™

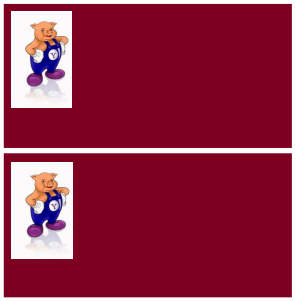

OOZIE
TM

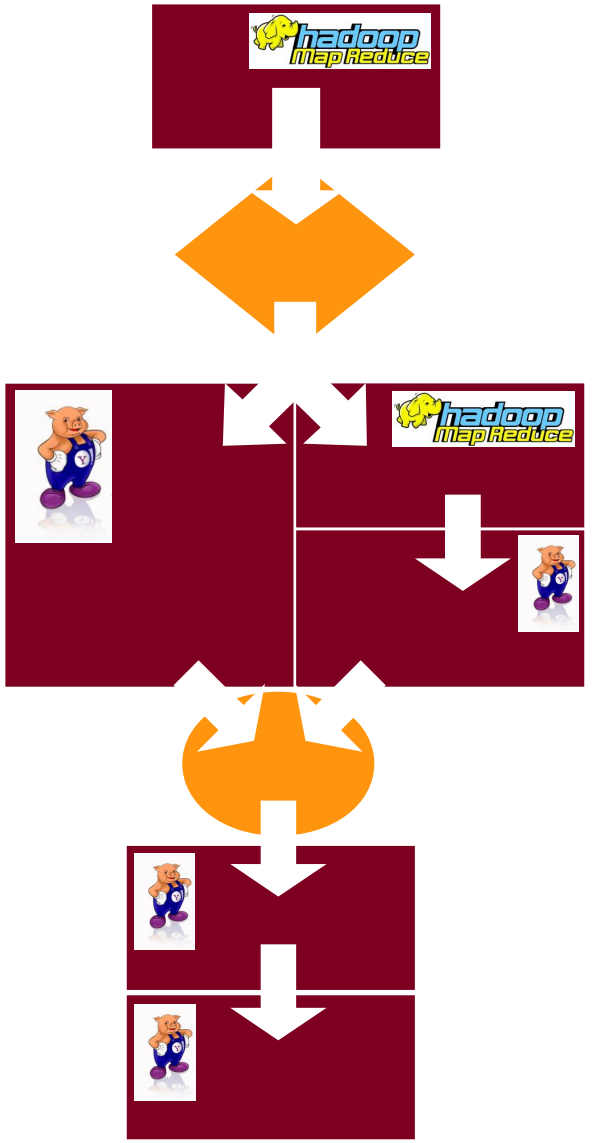

hadoop
Map Reduce



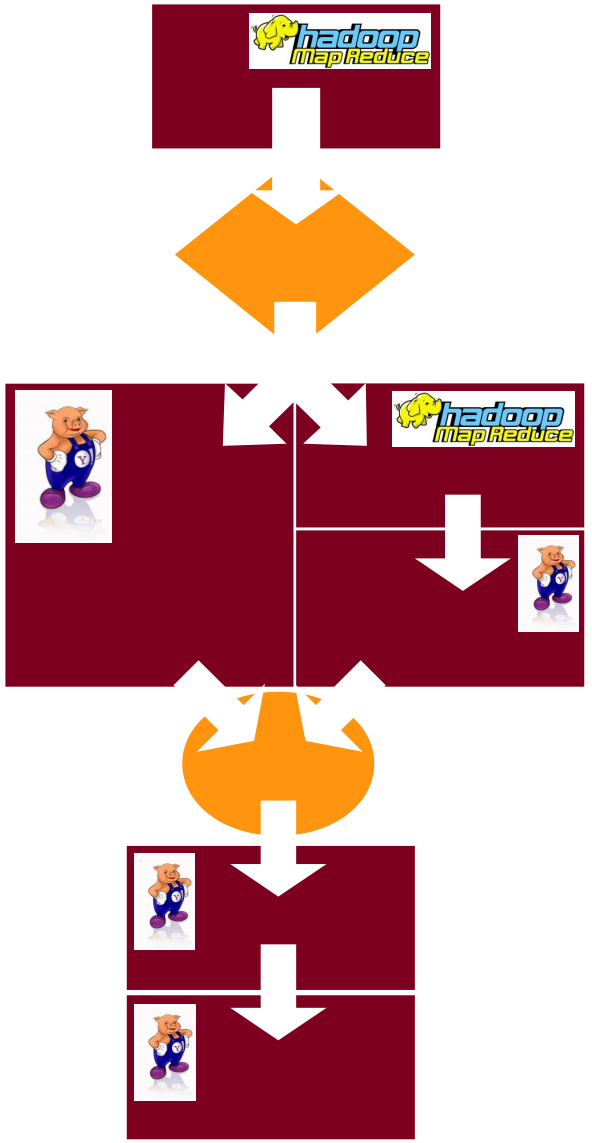

hadoop
Map Reduce


hadoop
Map Reduce





OOZIE™



Trigger the workflow:

Manually
On data arrival
Based on a schedule

Avoid job conflicts.

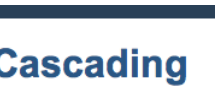
Java API
Pipes C++ API
Streaming
Dumbo for Python
...

TM

TM

TM

TM



TM

TM

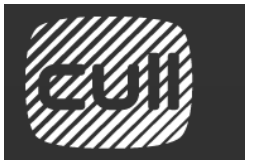
TM

About MRUnit

MRUnit is a Java library that helps devel


TM
mahout

Show users content they like



TM

Show users content they like

SpeedDate™

CUII

iOffer
A Place to Buy, Sell & Trade
Clothing & Accessories

mippin
the internet, mobile

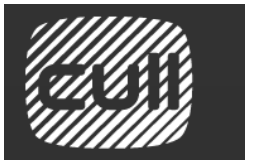
Drupal™

Monitor media and social media

365MEDIA

NewsCred If the news matters, we have it
HOME CONTENT BUYERS

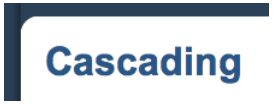
Show users content they like



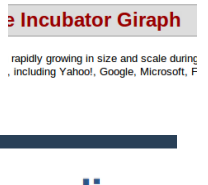
Monitor media and social media

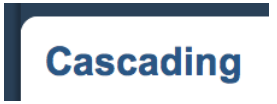
Deliver ads targeted to users

Java API
Pipes C++ API
Streaming
Dumbo for Python
...



Java API
Pipes C++ API
Streaming
Dumbo for Python
...





Export much smaller result set to database and use standard tools.

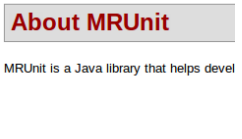
Use custom visualization.

TM

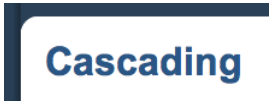
TM

TM

Java API
Pipes C++ API
Streaming
Dumbo for Python
...



TM



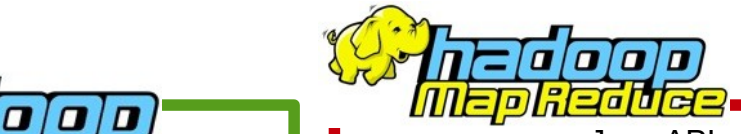
TM

TM

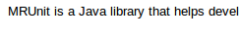
TM

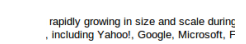
Export much smaller result set to database and use standard tools.

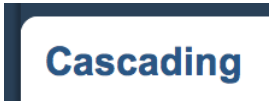
Use custom visualization.



Java API
Pipes C++ API
Streaming
Dumbo for Python
...

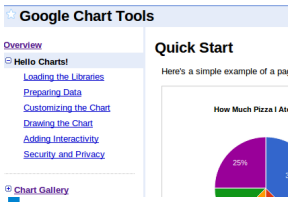


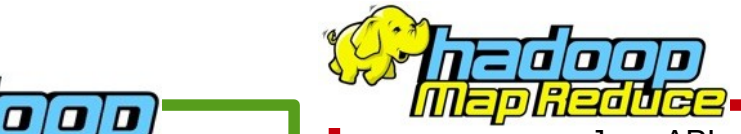




Export much smaller result set to database and use standard tools.

Use custom visualization.

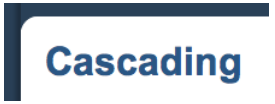




Java API
Pipes C++ API
Streaming
Dumbo for Python
...

MRUnit is a Java library that helps devel

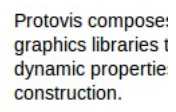
rapidly growing in size and scale during
, including Yahoo!, Google, Microsoft, F



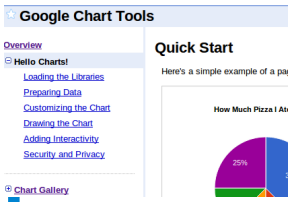
Export much smaller result set to database and use standard tools.

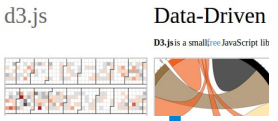
Use custom visualization.

Export to file format compatible with your favourite graphing tool.



Protovis compose
graphics libraries t
dynamic propertie
construction.



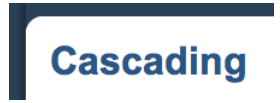


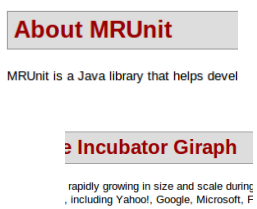
Java API
Pipes C++ API
Streaming
Dumbo for Python
...

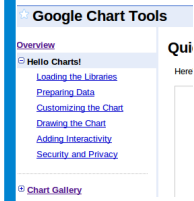
Export m
result set
and use

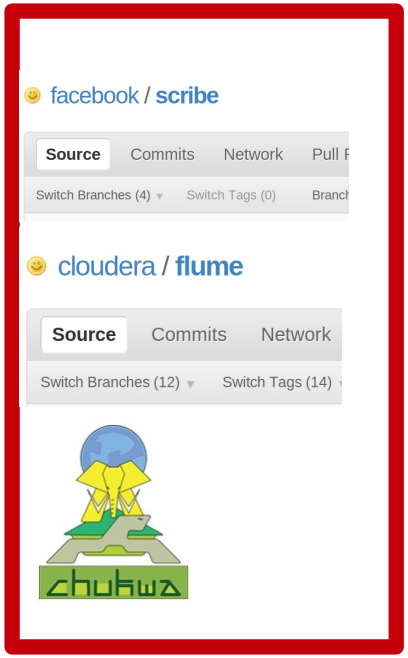
Use cust

Export to
compatib
favourite





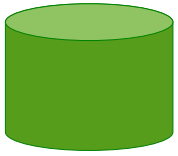


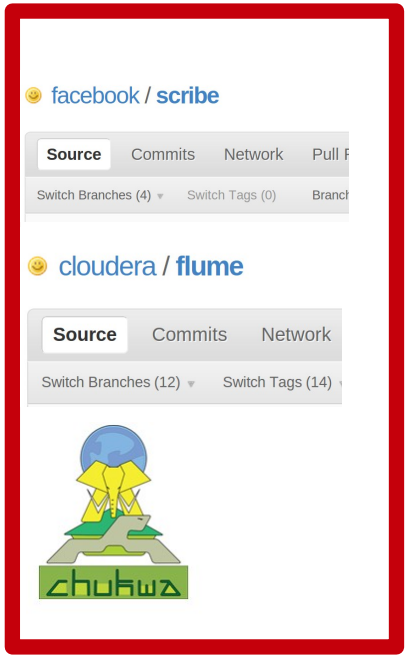

facebook / scribe
Source
Commits
Network
Switch Branches (4)
Switch Tags (0)
cloudera / flume
Source
Commits
Network
Switch Branches (12)
Switch Tags (14)
chukwa


hadoop
HDFS
Google code protobuf
Protocol Buffers - Google's data interchange format
AVRO
TM
Thrift
TM
HBASE
TM
HYPERTABLE
HIVE
TM


hadoop
Map Reduce
Java API
Pipes C++ API
Streaming
Dumbo for Python
...
TM
About MRUnit
MRUnit is a Java library that helps devel
Incubator Giraph
rapidly growing in size and scale during
, including Yahoo!, Google, Microsoft, F
Cascading
mahout
TM
OOZIE
TM

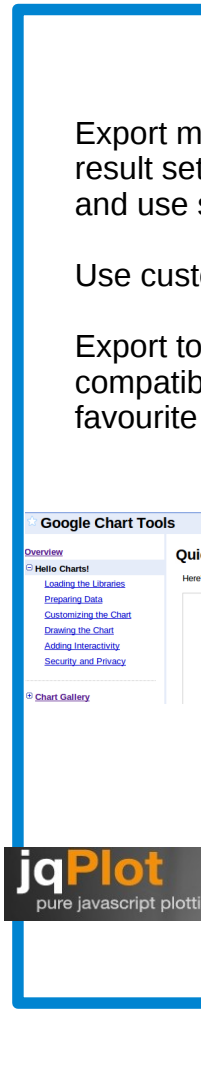

Export m
result set
and use
Use cust
Export to
compatib
favourite
Google Chart Tools
Overview
Hello Charts!
Loading the Libraries
Preparing Data
Customizing the Chart
Drawing the Chart
Adding Interactivity
Security and Privacy
Chart Gallery
jqPlot
pure javascript

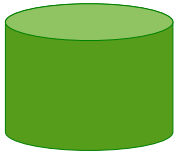


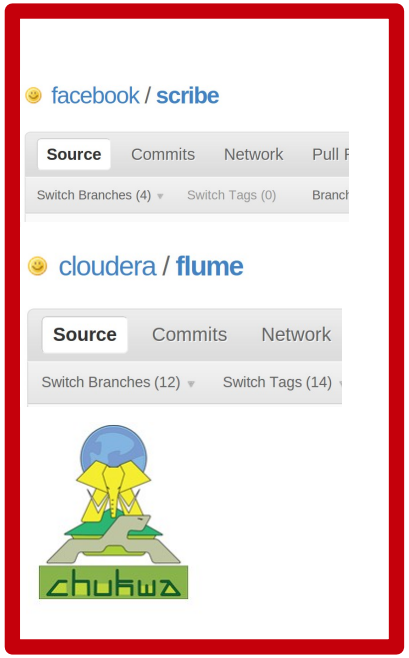

facebook / scribe
Source
Commits
Network
Switch Branches (4)
Switch Tags (0)
cloudera / flume
Source
Commits
Network
Switch Branches (12)
Switch Tags (14)
chukwa

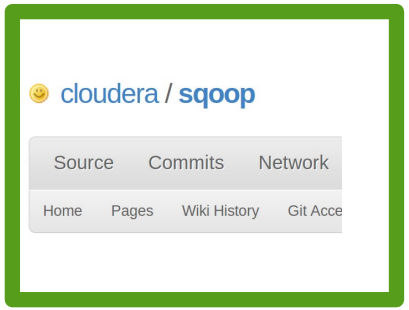

hadoop
HDFS
Google code protobuf
Protocol Buffers - Google's data interchange format
TM
AVRO
TM
Thrift
TM
HBASE
HYPERTABLE
HIVE
TM


hadoop
Map Reduce
Java API
Pipes C++ API
Streaming
Dumbo for Python
...
TM
About MRUnit
MRUnit is a Java library that helps devel
Incubator Giraph
rapidly growing in size and scale during
, including Yahoo!, Google, Microsoft, F
Cascading
TM
mahout
TM
OOZIE


Export m
result set
and use
Use cust
Export to
compatib
favourite
Google Chart Tools
Overview
Hello Charts!
Loading the Libraries
Preparing Data
Customizing the Chart
Drawing the Chart
Adding Interactivity
Security and Privacy
Chart Gallery
jqPlot
pure javascript plotti

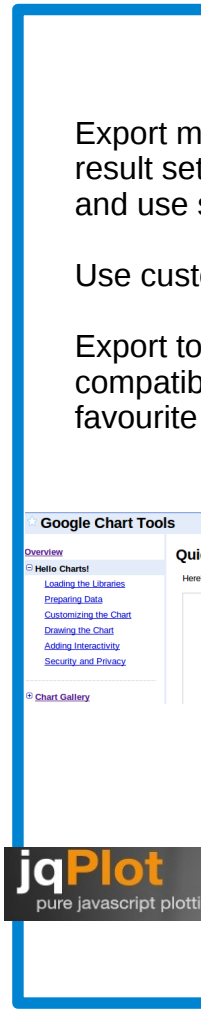



facebook / scribe
Source
Commits
Network
Switch Branches (4)
Switch Tags (0)
cloudera / flume
Source
Commits
Network
Switch Branches (12)
Switch Tags (14)
chukwa


cloudera / sqoop
Source
Commits
Network
Home
Pages
Wiki History
Git Acce


hadoop
HDFS
Google code protobuf
Protocol Buffers - Google's data interchange format
AVRO
TM
Thrift
TM
HBASE
TM
HYPERTABLE
HIVE
TM


hadoop
Map Reduce
Java API
Pipes C++ API
Streaming
Dumbo for Python
...
TM
About MRUnit
MRUnit is a Java library that helps devel
Incubator Giraph
rapidly growing in size and scale during
, including Yahoo!, Google, Microsoft, F
Cascading
mahout
TM
OOZIE
TM


Export m
result set
and use
Use cust
Export to
compatib
favourite
Google Chart Tools
Overview
Hello Charts!
Loading the Libraries
Preparing Data
Customizing the Chart
Drawing the Chart
Adding Interactivity
Security and Privacy
Chart Gallery
jqPlot
pure javascript plotti

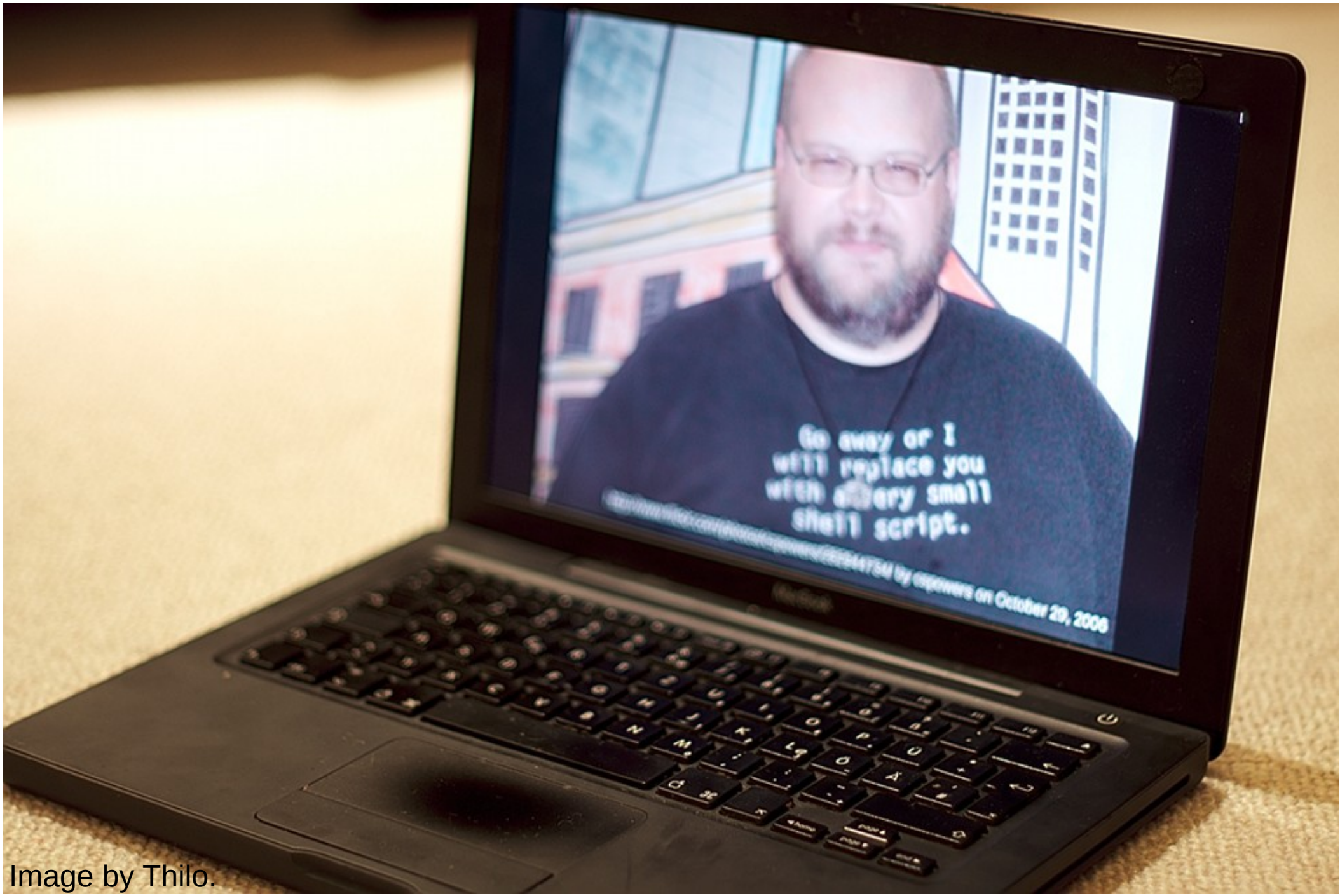
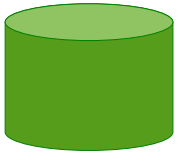


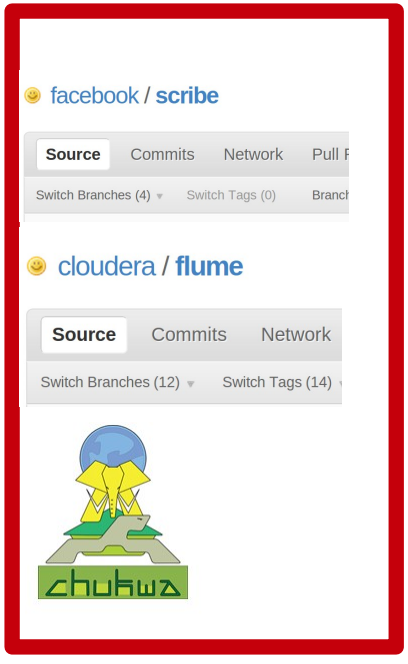
Image by Thilo.

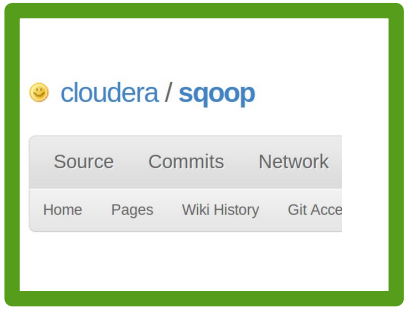
(Thanks to Thilo for helping set up the cluster, Thanks to packet and masq for two of the three machines.)

```
package { 'openssh-server':
    ensure => installed,
}

file { '/etc/ssh/sshd_config':
    source  => 'puppet:///modules/sshd/
sshd_config',
    owner   => 'root',
    group   => 'root',
    mode    => '640',
    notify  => Service['sshd'], # sshd
               will restart whenever you
               edit this file.
    require => Package['openssh-server'],
}

service { 'sshd':
    ensure => running,
    enable => true,
    hasstatus => true,
    hasrestart => true,
}
```

- Managing:
  - local files
  - packages
  - Services
- Configuring:
  - cron
  - users and groups

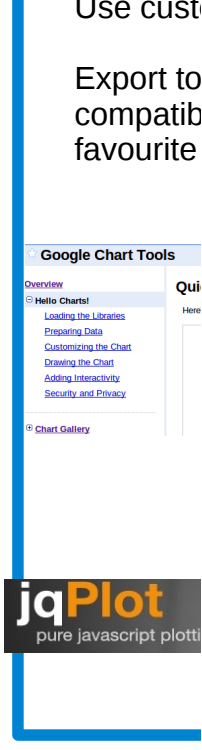



facebook / scribe
Source
Commits
Network
Switch Branches (4)
Switch Tags (0)
cloudera / flume
Source
Commits
Network
Switch Branches (12)
Switch Tags (14)
chukwa


cloudera / sqoop
Source
Commits
Network
Home
Pages
Wiki History
Git Acce


hadoop
HDFS
Google code protobuf
Protocol Buffers - Google's data interchange format
AVRO
TM
Thrift
TM
HBASE
TM
HYPERTABLE
HIVE
TM


hadoop
Map Reduce
Java API
Pipes C++ API
Streaming
Dumbo for Python
...
TM
About MRUnit
MRUnit is a Java library that helps devel
Incubator Giraph
rapidly growing in size and scale during
, including Yahoo!, Google, Microsoft, F
Cascading
mahout
TM
OOZIE
TM

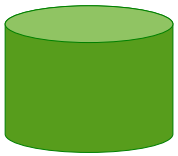

Export m
result set
and use
Use cust
Export to
compatib
favourite
Google Chart Tools
Overview
Hello Charts!
Loading the Libraries
Preparing Data
Customizing the Chart
Drawing the Chart
Adding Interactivity
Security and Privacy
Chart Gallery
jqPlot
pure javascript


Chef

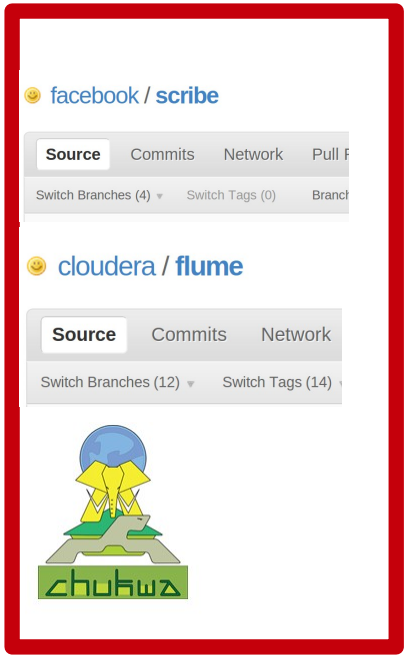

puppet labs
How to Buy
Documentation
Products
Introducing Puppet

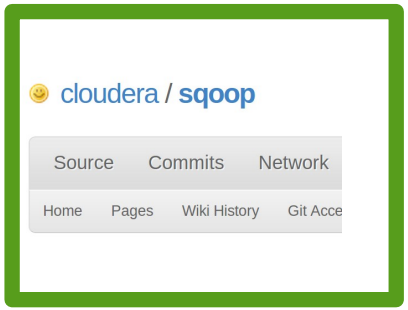
TM

- Started as distributed synchronization.
- Configuration management.
- Provides for distributed naming.
- Group services.



TM





TM

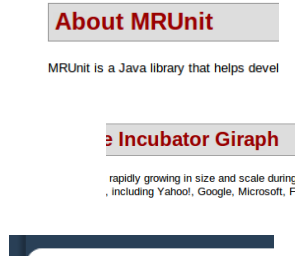
TM

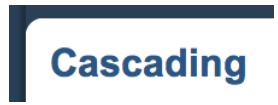
TM

TM

Java API
Pipes C++ API
Streaming
Dumbo for Python
...

TM





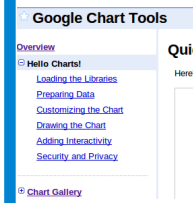
TM

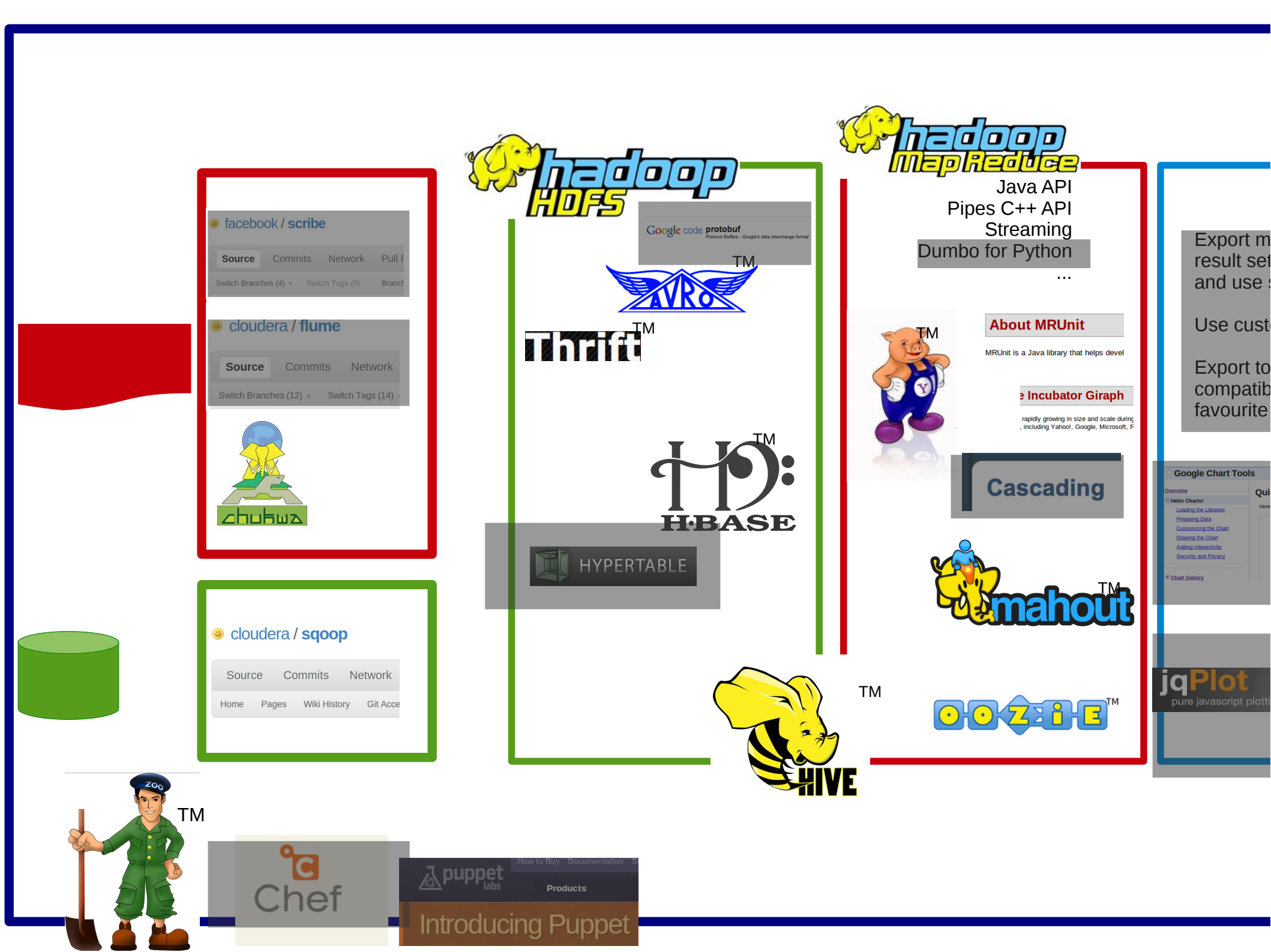
TM

Export m
result set
and use

Use custo

Export to
compatib
favourite

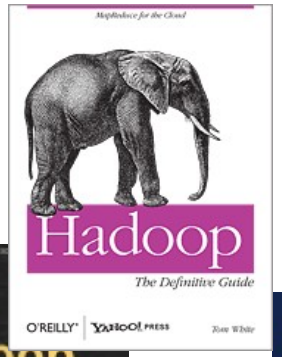


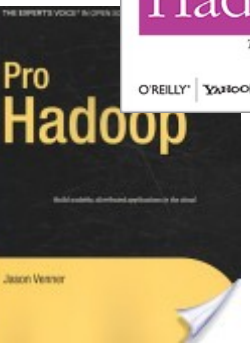

hadoop HDFS
hadoop Map Reduce
Java API
Pipes C++ API
Streaming
Dumbo for Python
...
facebook / scribe
Source
Commits
Network
cloudera / flume
Source
Commits
Network
Switch Branches (12)
Switch Tags (14)
chukwa
cloudera / sqoop
Source
Commits
Network
Home
Pages
Wiki History
Git Acce
Google code protobuf
Protocol Buffers - Google's data interchange format
TM
AVRO
TM
Thrift
TM
HBASE
HYPERTABLE
TM
About MRUnit
MRUnit is a Java library that helps devel
Incubator Giraph
rapidly growing in size and scale during
, including Yahoo!, Google, Microsoft, F
Cascading
mahout
TM
HIVE
TM
OOZIE
TM
Export m
result set
and use
Use cust
Export to
compatib
favourite
Google Chart Tools
Overview
Hello Charts!
Loading the Libraries
Preparing Data
Customizing the Chart
Drawing the Chart
Adding Interactivity
Security and Privacy
Chart Gallery
jqPlot
pure javascript plotti
ZOO
TM
Chef
puppet labs
How to Buy
Documentation
Products
Introducing Puppet

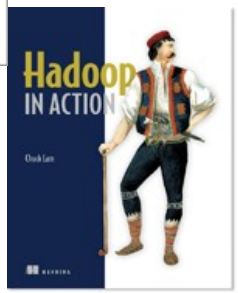
Jumpstart your project with proven code.

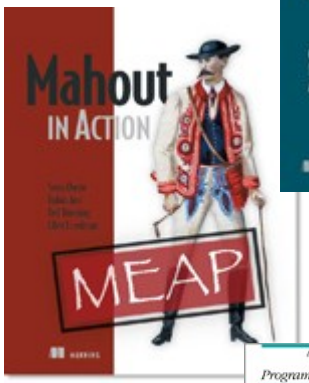
January 8, 2008 by dreizehn28
http://www.flickr.com/photos/1328/2176949559

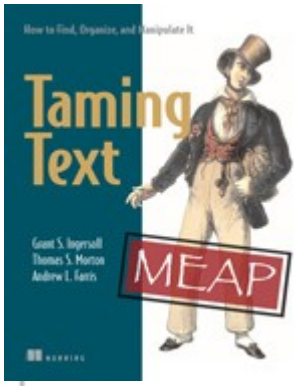
There is help

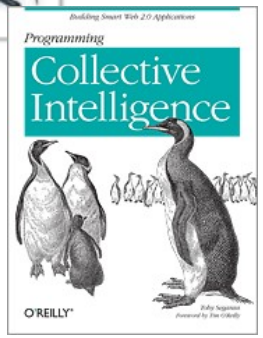


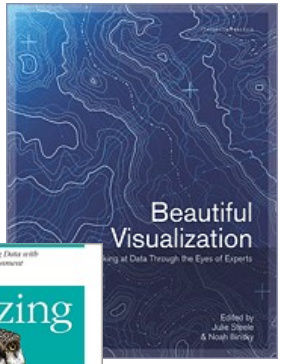


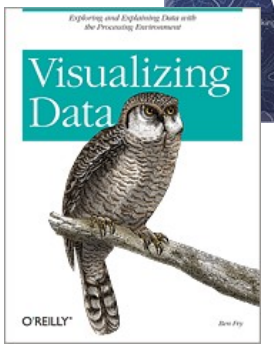


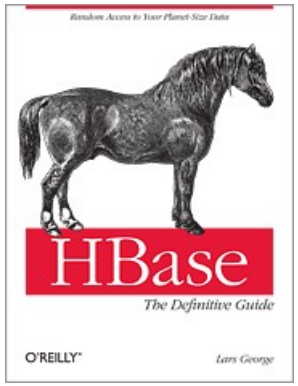


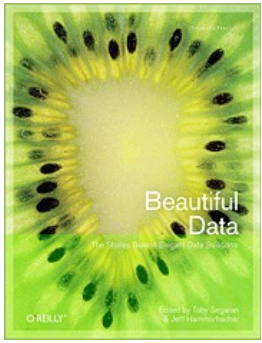












There is help

- Cloudera
- Hortonworks
- MapR
- JTeam
- 101tec
- Datameer
- Sematext
- ...


Discuss ideas and problems online.
November 16, 2005 [phil h]
http://www.flickr.com/photos/hi-phi/64055296

WARNING

Image by Shane Fullwood: http://www.flickr.com/photos/whassupbud55/4581192168/

http://flickr.com/photos/iamdietrich/4075363845/
By: Iam Dietrich

May 1, 2007 by danny angus
http://www.flickr.com/photos/killerbees/479864437/

user@.apache.org

dev@.apache.org

Image by: Patrick McEvoy

Interest in solving hard problems.

Being part of lively community.

Engineering best practices.

Bug reports, patches, features.

Documentation, code, examples.