



推动云计算应用： Hadoop 开源平台

郑皓
首席架构师
雅虎全球软件研究和开发中心 北京



Agenda

- 雅虎Yahoo!公司的云计算战略
- Hadoop 开源社区和用户团体
- Hadoop 和 Yahoo!
- Hadoop 在 Yahoo! 的应用

为什么云计算对Yahoo!至关重要

我们的使命：成为亿万网民在线活动的中心



HUNDREDS

资产 / 平台

600M

独立用户 / 月

300M+

MAIL 用户 / 月

HUNDREDS

PB量级存储

BILLIONS

对象存储

PETABYTES

每日流量



Yahoo!如何使用云计算平台

- 依赖云计算来提升用户和广告主的体验：
 - 1) 处理和分析海量数据
 - 2) 在全球范围内加速内容投放
 - 3) 灵活存储
- 大规模促进创新和科学的研究
- 和开源社区合作共享我们的云计算平台

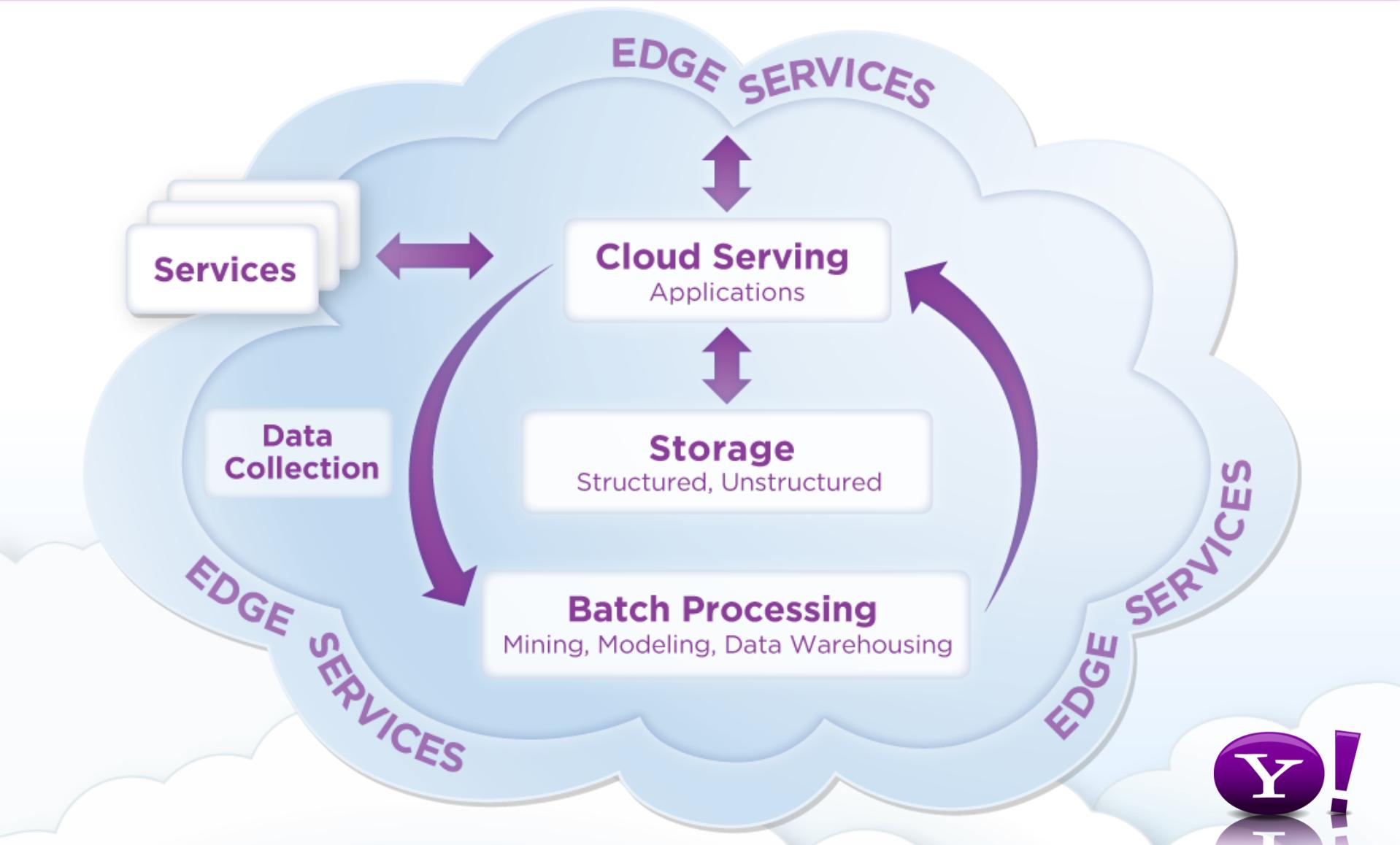


Yahoo!的云战略

- 创建Yahoo!自己的私有云
- 优化Yahoo!全球资产
- 数据处理和服务环境
- 核心收益：驱动创新
- 成熟之后将核心技术开源
- 付出了多年努力



Yahoo! 云内部架构





业界挑战：日积月累的海量数据

1

大量优质数据积累在相对
少数人身上

2

从大量数据中挖掘有用信
息依然一个挑战

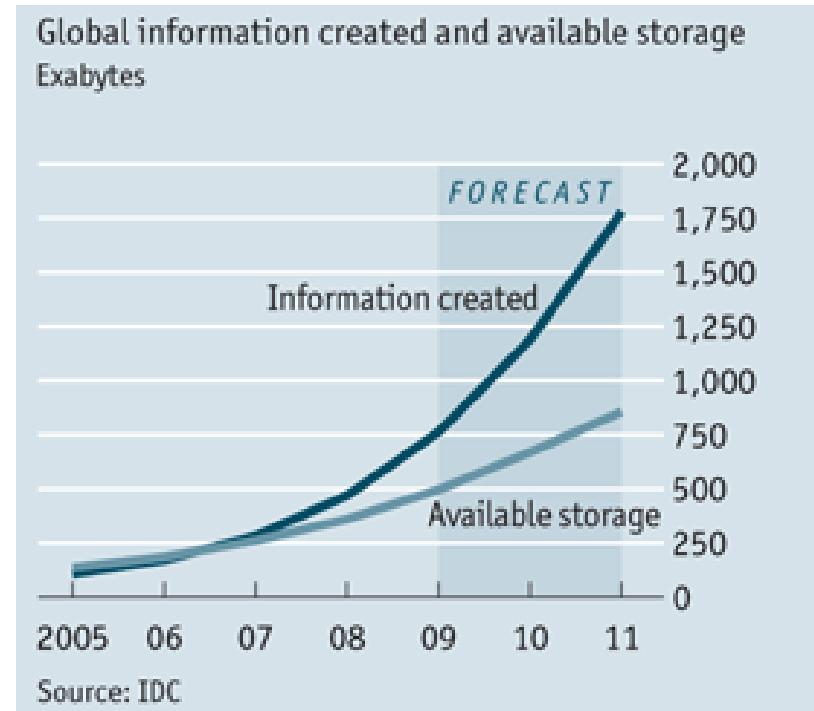
3

数据成为竞争力的分水岭：
搜索日志，广告点击数据，
浏览历史，社交图，等等

Yahoo!的技术解决方案：
Hadoop

“现在是数据的工业革命。” - Joe
Hellerstein, 加州伯可莱大学

“到2020年，全球数据规模会比2009年扩
大44倍。” [IDC]



YAHOO!



Hadoop是什么?

- 一个分布式文件系统和并行执行环境
- 让用户便捷地处理海量数据
- Apache软件基金会下面的一个开源项目
- 目前Yahoo!是最主要的贡献者



快速增长的用户群

2007

YAHOO!



last.fm

2008

Google™

ImageShack
online media hosting

Lookery

able grape

A9™
facebook

Cascading

IBM®

enORMO

krugle

INFORMATION
SCIENCES
INSTITUTE



zvents

Discover Things To Do



The New York Times



Cornell University
Computing and Information Science

News Corporation

LOTAME
parc®
Palo Alto Research Center

Visible
MEASURES

veoh™

rockspace®
HOSTING

2009 - 2010

AOL

WorldLingo

cloudera

cooliris

deepdyve
BETA

TEXTMAP
THE ENTITY SEARCH ENGINE



PSG College of Technology
The Technology that leads the world

eyealike
slike

iterend

tailsweep
adknowledge

hulu™

RapLeaf

Ning

quantcast

USCIMS

pressflip

HOSTING
HABITAT

detikSearch



stampede
beta

VK SOLUTIONS
Global Solutions Provider

Microsoft

Terrier



CbIR

HOLA
SOURCES

Systems@ETH
Zürich

NETFLIX

amazon
web services™

twitter™

LinkedIn

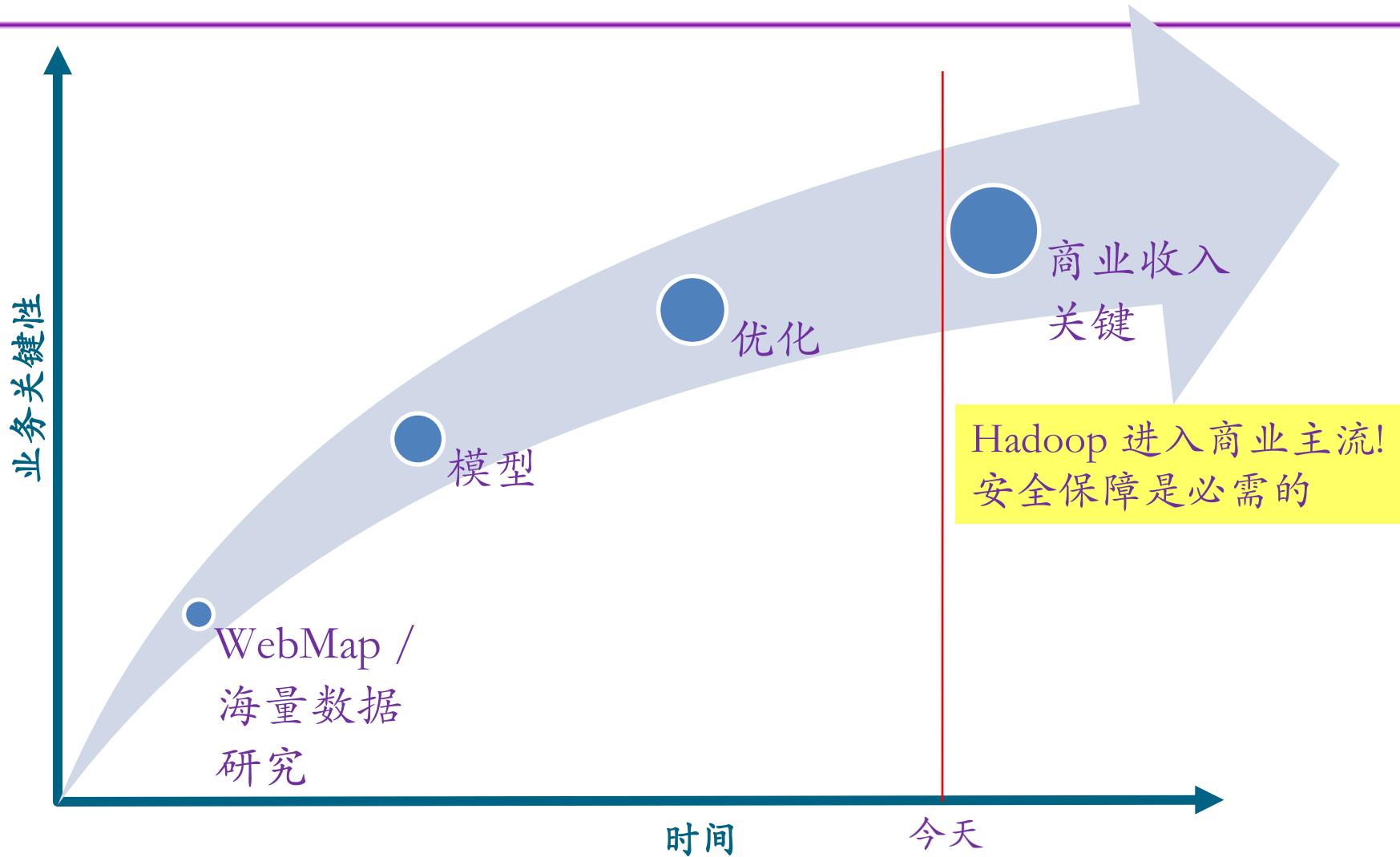
eBay®

TARAGANA
Innovation • Quality • Simplicity

YAHOO!



Hadoop逐渐成为主流云计算平台





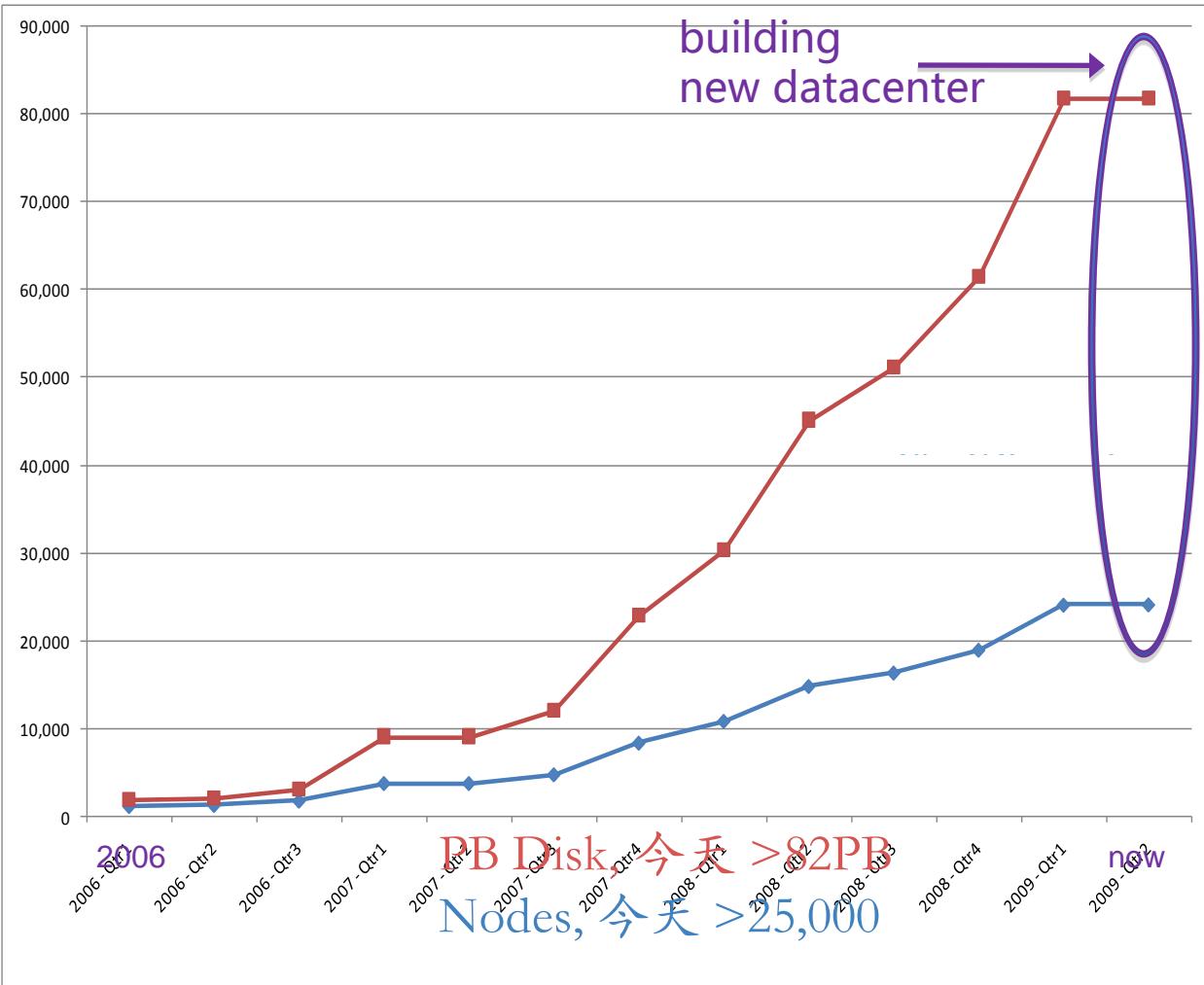
Yahoo!和Hadoop密切关系

- Yahoo!是：
 - Hadoop最大的用户
 - Hadoop最大的测试者
 - Hadoop最大的贡献者
- 还有：
 - 我们发布了“*Yahoo! Hadoop 公开版*”
 - 我们贡献了所有在Hadoop上的工作给Apache软件基金会
 - 我们持续积极地投入到Hadoop开发中
 - 我们不以Hadoop的服务和支持盈利！
 - 我们使用Hadoop来驱动整个Yahoo!

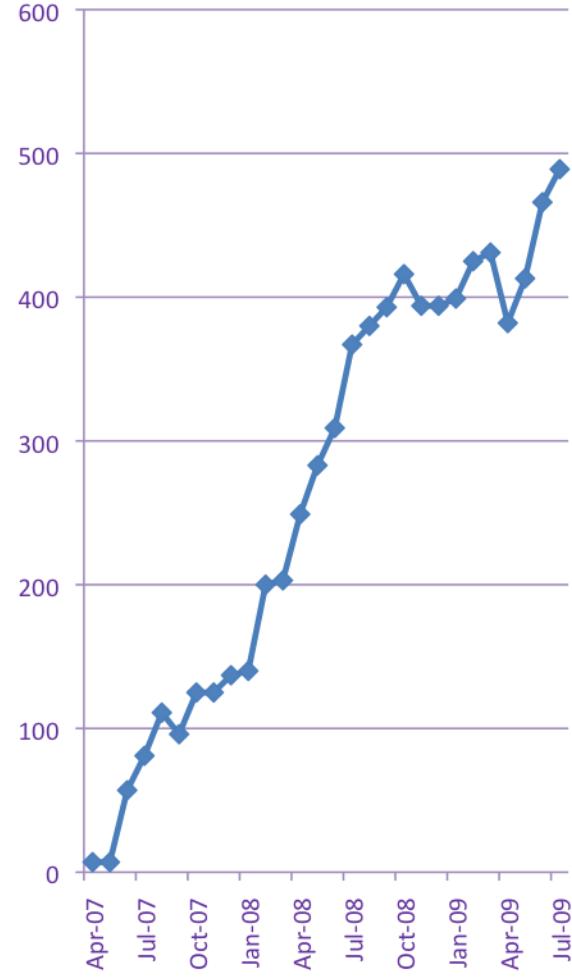


Hadoop最大的用户

硬件规模



内部Hadoop用户数



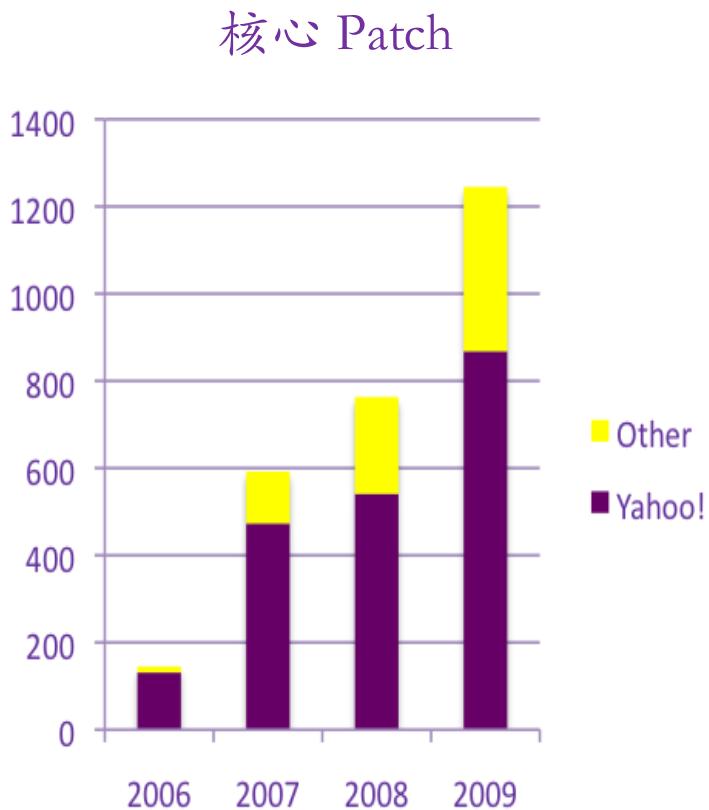


Hadoop最大的测试者

- *Yahoo! Hadoop*公开放的每一次发布都经过了很多级别的严格测试
- Hadoop集群的4个层级：
 - 开发, 测试, 和质检 (~10%的硬件)
 - 持续集成 / 测试最新代码
 - 概念证明和特设工作 (~10%的硬件)
 - 运行最新版本
 - 科学与研究 (~60%的硬件)
 - 运行比较稳定的版本
 - 产品 (~20%的硬件)
 - 运行最稳定的版本



Hadoop最大的贡献者



- 主要的patch都来自Yahoo!
 - 72% 的核心 patch
 - Core = HDFS, Map-reduce, Common
 - 存在低估现象
 - 一些雅虎员工使用apache.org的账户
 - 全职contributors提交大的patch
- 截止目前，Yahoo!是Hadoop贡献者的最大雇主
 - 我们将在Hadoop上做的所有工作回馈给Apache社区
 - 团队: 美国, 班加罗尔, 北京
 - <http://careers.yahoo.com>

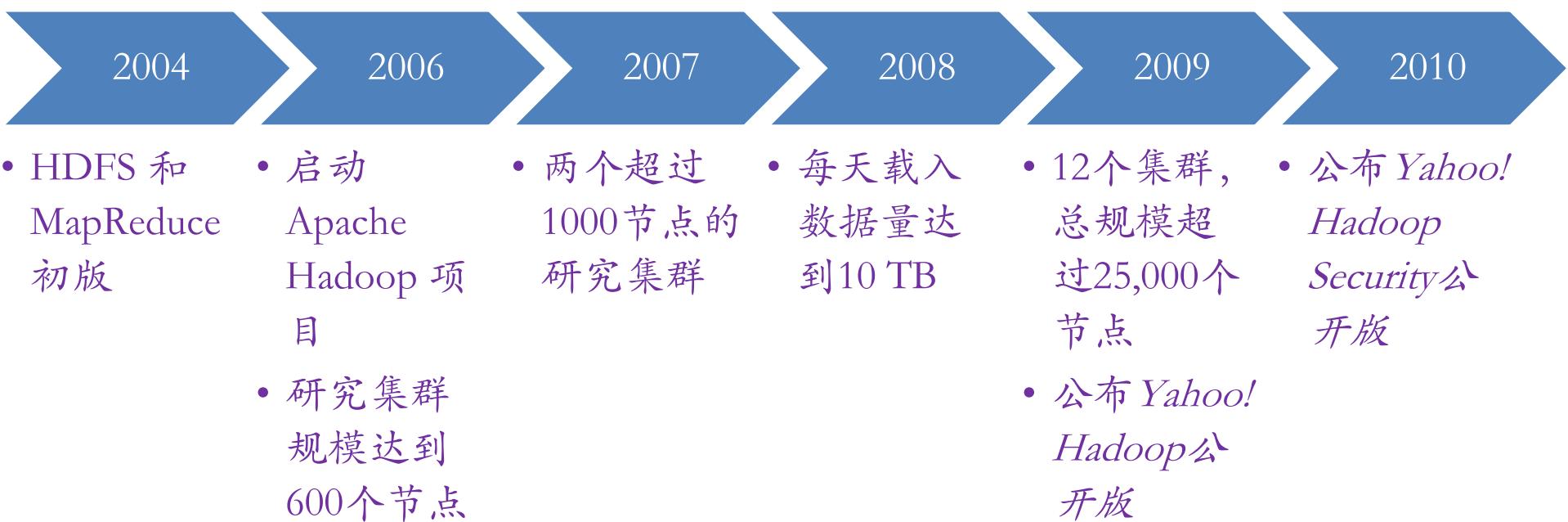


缘何 Yahoo! 选择 Hadoop

- 数据规模
 - 每月超过6亿的独立用户
 - 每天产生数十亿的transaction
 - PB级别的数据
- 分析和处理数据非常关键
 - 需要及时处理所有数据
 - 大量调研以寻找更好的模式，分析数据以生成各种报表
- 更低的成本需求
 - 使用低成本的通用硬件
 - 多项目之间共享资源
 - 在大规模集群上快速完成新的实验
 - 每天需要处理许多硬件故障
- Hadoop这一基础架构可以满足这些需求

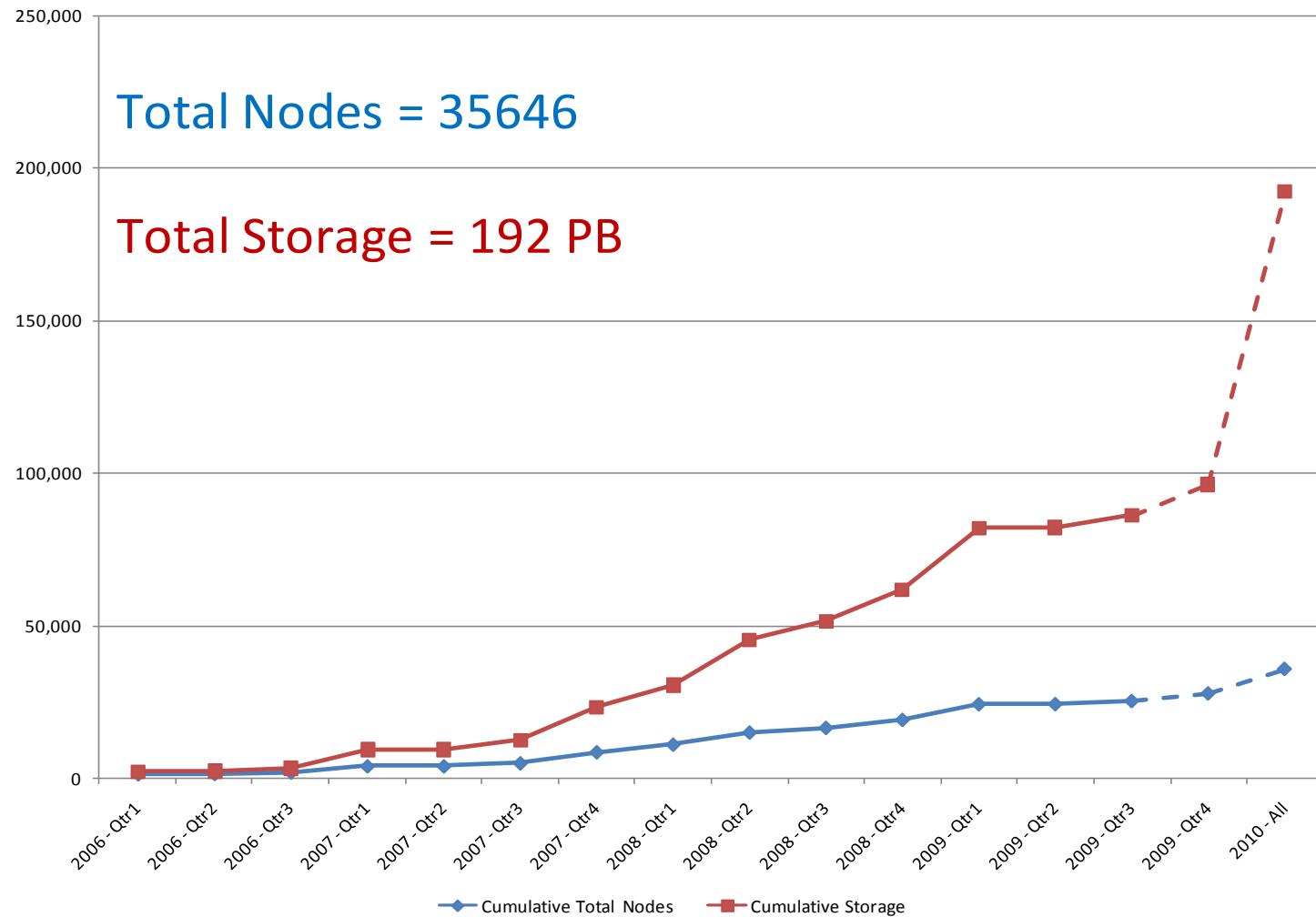


Yahoo! 使用 Hadoop 时间表





Yahoo!使用Hadoop趋势图





Yahoo! Front Page – 案例分析

Y! View Yahoo! Sites ▾

MY FAVORITES [+ Add](#)

- [Yahoo! Mail](#)
- [Autos](#)
- [eBay](#)
- [Finance \(Dow Jones\)](#)
- [Flickr](#)
- [Games](#)
- [Horoscopes](#)
- [Maps](#)
- [Messenger](#)
- [Movies](#)
- [Music](#)
- [MySpace](#)
- [Personals](#)
- [Sports](#)
- [Weather \(65°F\)](#)

RECOMMENDED

- [Deal Of The Day](#)
- [Buzz](#)
- [Shine](#)

[Edit](#) [+ Add](#)



Man, woman with same name to wed

Kelly Hildebrandt is in love with Kelly Hildebrandt, and soon they'll be married. [» How they met](#) NBC Miami

- Santa's helpers marry
- Unusual proposals
- Yahoo! Buzz

[Unique name brings love](#) [Eclipse spooks superstitious](#) [Fighter's cool pool jump](#) [Crowe pulls a 'Robin Hood'](#)

[« Prev](#) [•](#) [•](#) [•](#) [•](#) [•](#) [Next »](#)

[NEWS](#) [WORLD](#) [LOCAL](#) [FINANCE](#)

- Obama pushes back against critics of health care overhaul
- Economic indicators up more than expected in June
- July becomes deadliest month for U.S. in Afghanistan
- Clinton: Officials believe 9/11 ringleaders are hiding in...
- Spacewalk unfolds on 40th moon landing anniversary
-  Company designs suit to simulate elderly driving
- Sony bids \$50 million for rights to Jackson rehearsal film
- Defense Secretary Gates announces Army being... - SJ Mercury...
- Frank McCourt - author of 'Angela's Ashes' - S.F. Chronicle
- Annual all-star football game a family affair - Sunnyvale Sun

updated 12:04 pm PDT

More: [News](#) [Popular](#) [Buzz](#)

Markets: Dow: 8,821.70 0.88% Nasdaq: 1,900.55 0.73%

Enter stock symbol [Get Quotes](#)

POPULAR SEARCHES

1. Paula Abdul
2. Katherine Heigl
3. Angela's Ashes
4. Jon Gosselin
5. David Beckham
6. Lupus
7. Moon Landing
8. Tour de France
9. Pearl Jam
10. George Clooney

PROGRESSIVE[®] DIRECT

**The Name Your Price[®] option.
New and only from Progressive.**



Enter ZIP Code: [Name Your Price](#)

[Compare and Save](#) - Ad Feedback

SPOTLIGHT

[« Prev](#) [Next »](#)

Hot list: Must-watch videos

- 
- Guinea pigs love watermelon
 - How Ron Paul got 'punk'd' by 'Bruno'
 - Michael Jackson flash mob
 - Inside Vogue: The September Issue'

[Dog vs. playground slide »](#)



GEICO Car Insurance
You could save over \$500 on car insurance.
Get a free quote today.

YAHOO!



Yahoo! Front Page – 案例分析

Y! View Yahoo! Sites ▾

MY FAVORITES [+ Add](#)

- [Yahoo! Mail](#)
- [Autos](#)
- [eBay](#)
- [Finance \(Dow Jones\)](#)
- [Flickr](#)
- [Games](#)
- [Horoscopes](#)
- [Maps](#)
- [Messenger](#)
- [Movies](#)
- [Music](#)
- [MySpace](#)
- [Personals](#)
- [Sports](#)
- [Weather \(65°F\)](#)

RECOMMENDED

- [Deal Of The Day](#)
- [Buzz](#)
- [Shine](#)

[Edit](#) [+ Add](#)



Man, woman with same name to wed

Kelly Hildebrandt is in love with Kelly Hildebrandt, and soon they'll be married. [» How they met](#) NBC Miami

- Santa's helpers marry
- Unusual proposals
- Yahoo! Buzz

[Unique name brings love](#) [Eclipse spooks superstitious](#) [Fighter's cool pool jump](#) [Crowe pulls a 'Robin Hood'](#)

[« Prev](#) [•](#) [•](#) [•](#) [•](#) [•](#) [Next »](#)

[NEWS](#) [WORLD](#) [LOCAL](#) [FINANCE](#)

- Obama pushes back against critics of health care overhaul
- Economic indicators up more than expected in June
- July becomes deadliest month for U.S. in Afghanistan
- Clinton: Officials believe 9/11 ringleaders are hiding in...
- Spacewalk unfolds on 40th moon landing anniversary
-  Company designs suit to simulate elderly driving
- Sony bids \$50 million for rights to Jackson rehearsal film
- Defense Secretary Gates announces Army being... - SJ Mercury...
- Frank McCourt - author of 'Angela's Ashes' - S.F. Chronicle
- Annual all-star football game a family affair - Sunnyvale Sun

updated 12:04 pm PDT

More: [News](#) [Popular](#) [Buzz](#)

Markets: Dow: 8,821.70 0.88% Nasdaq: 1,900.55 0.73%

Enter stock symbol [Get Quotes](#)

POPULAR SEARCHES

1. Paula Abdul
2. Angela's Ashes
3. David Beckham
4. Moon Landing
5. Pearl Jam
6. Katherine Heigl
7. Joe Cosselin
8. Lupus
9. Tour de France
10. George Clooney

檢索索引

PROGRESSIVE® DIRECT

The Name Your Price® option.
New and only progressive.



Enter ZIP Code: Name Your Price [Compare and Save](#) - Ad Feedback

SPOTLIGHT

[« Prev](#) [Next »](#)

Hot list: Must-watch videos

- Guinea pigs love watermelon
- How Ron Paul got 'punk'd' by 'Bruno'
- Michael Jackson flash mob
- Inside Vogue: The September Issue'

[Dog vs. playground slide »](#)

GEICO Car Insurance
You could save over \$500 on car insurance.
Get a free quote today.

YAHOO!



Yahoo! Front Page – 案例分析

Y! View Yahoo! Sites ▶

MY FAVORITES [+ Add](#)

- [Yahoo! Mail](#)
- [Autos](#)
- [eBay](#)
- [Finance \(Dow Jones ↑\)](#)
- [Flickr](#)
- [Games](#)
- [Horoscopes](#)
- [Maps](#)
- [Messenger](#)
- [Movies](#)
- [Music](#)
- [MySpace](#)
- [Personals](#)
- [Sports](#)
- [Weather \(65°F\)](#)

RECOMMENDED

- [Deal Of The Day](#)
- [Buzz](#)
- [Shine](#)

[Edit](#) [+ Add](#)

YAHOO! MAIL Jonathandoe58 Available [Sign Out](#) [My Account](#) [Mail Classic](#)

Search the Web... [Search](#)

[Check Mail](#) [New](#) [Go](#)

[Delete](#) [Reply](#) [Forward](#) [Spam](#) [Move](#) [Print](#) [More Actions](#) [View](#)

Index 3 messages

	From	Subject	Date	Size
1	Jane Smith	Wedding pics	Tue, 6/17/08 11:37 AM	57KB
2	Andy Oswald	Re: Wedding	Tue, 6/17/08 10:26 AM	3KB
3	Jane Smith	Re: Wedding	Tue, 6/17/08 10:26 AM	111KB

[Download All 3 Images](#)

Hey Jon,
Here are some pictures from our beach wedding on the 14th. I hope you like them.
Jane Smith <jonathandoe58@yahoo.com>
Andy Oswald <andyoswald@yahoo.com>
Jane Smith <jonathandoe58@yahoo.com>

3 Images | [Download All 3 Images](#) | [Selected](#) | [Description](#)



- Economic indicators up more than expected in June
- July becomes deadliest month for U.S. in Afghanistan
- Clinton: Officials believe 9/11 ringleaders are hiding in...
- Spacewalk unfolds on 40th moon landing anniversary
-  Company designs suit to simulate elderly driving
- Sony bids \$50 million for rights to Jackson rehearsal film
- Defense Secretary Gates arrives in Iraq for first visit - SJ M...
- Frank McCourt, author of 'Angela's Ashes' - S.F. Chronicle
- Annual all-star football game a family affair - Sun-Times

内容管理

Updated 12:04 pm PDT [More: News | Popular | Buzz](#)

Markets: Dow: 8,821.70 **0.88%** Nasdaq: 1,900.55 **0.73%**

Enter stock symbol [Get Quotes](#)

POPULAR SEARCHES

- 1. Paula Abdul
 - 2. Angela's Ashes
 - 3. David Beckham
 - 4. Moon Landing
 - 5. Pearl Jam
 - 6. Katherine Heigl
 - 7. Jon Gosselin
 - 8. Lupi
 - 9. Tour de France
 - 10. George Clooney

An advertisement for Progressive Direct. A woman with long brown hair is smiling, wearing a white t-shirt with the 'PROGRESSIVE' logo and a blue and orange swoosh. She is holding a large, semi-transparent blue oval that covers the central text area. Inside the oval, the text 'The Name Your Price®' is at the top, followed by 'option.' and 'New and only' on the left, and 'from Progressive.' on the right. Below this, in large white characters, is the text '广告优化' (Advertisement Optimization).

SPOTLIGHT

«Prev | Next»

Hot list: Must-watch videos



Beamer background slide 8



[GEICO Car Insurance](#)
You could save over \$500 on car insurance
Get a free quote today.

YAHOO!



Search Assist™ – 案例分析



- 使用Hadoop构建数据库
- 3年的日志数据
- 20步的map-reduce计算

	Hadoop 前	Hadoop 后
耗时	26 天	20 分钟
语言	C++	Python
开始时间	2-3 星期	2-3 天



OpenCirrus™ 筋斗云

- OpenCirrus™ 筋斗云互助式云计算试验平台 成立于2009六月
- Yahoo!, HP, 和 Intel 联合发起
- 模拟一个现实的、全球化的互联网规模的环境
- 消除财务和后勤上的障碍
- 专注于数据密集型互联网计算的研究
- 给予研究人员前所未有的能力，来测试应用，测量基础设施以及构建于其上的服务的性能
- 促进产业界、学术界和政府部门的开放合作



问题?

郑皓

首席架构师

雅虎全球软件研究和开发中心 北京

信息咨询:

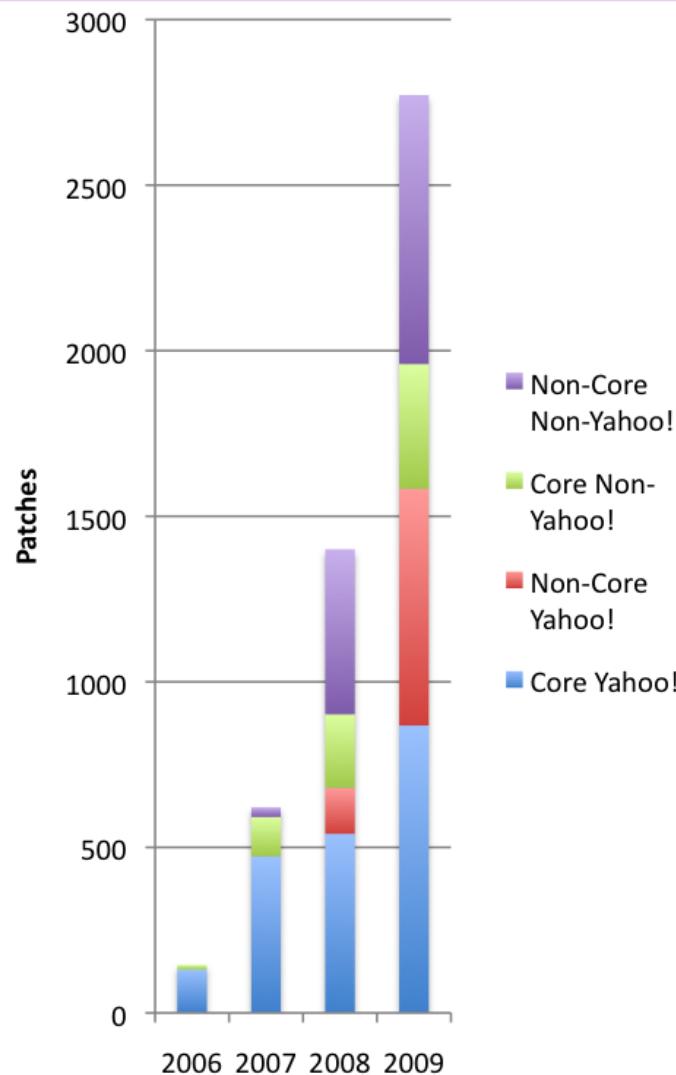
<http://hadoop.apache.org/>

<http://hadoop.yahoo.com/>

<http://beijing.yahoo.com/>



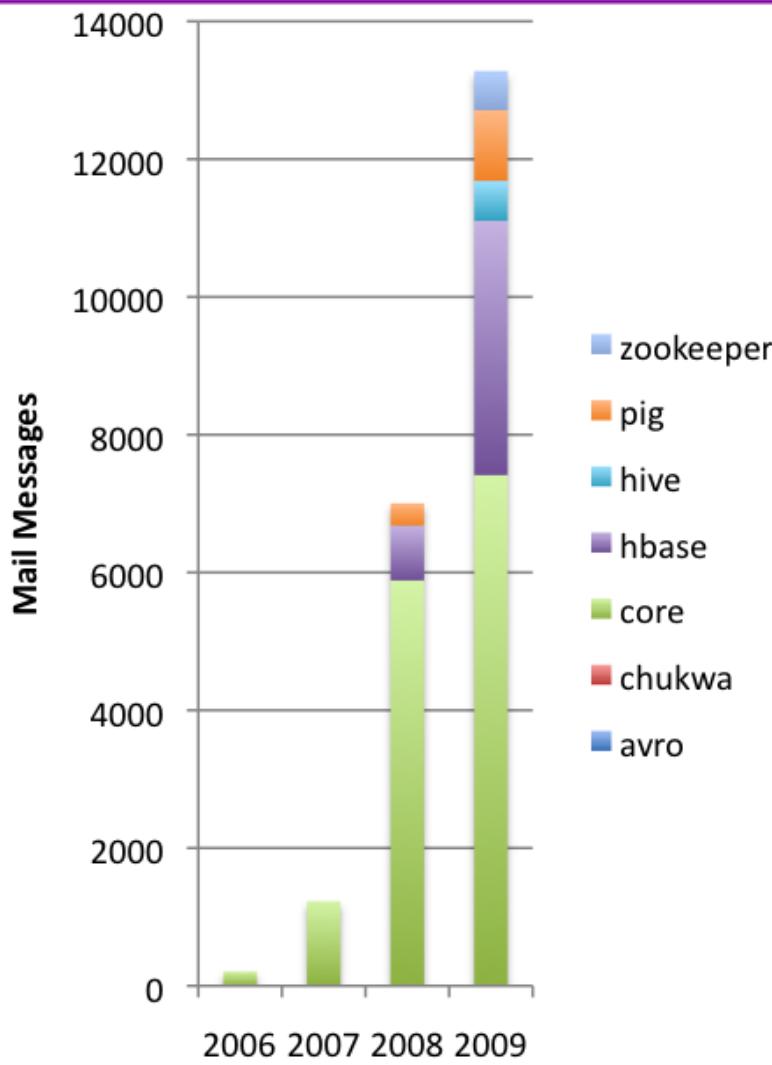
Patch 贡献年表



- 分为两个Sub-project
 - 核心 (HDFS & Map-Reduce)
 - 其他
- 所有分类都在增长
- 在2009 非核心>核心!
- 核心贡献者
 - 185 人 (30% 来自雅虎)
 - 72%的patch来自雅虎



子项目的用户列表



- 用户邮件量
 - 使用好的指标
 - 社团直接的衡量
 - 6个月旧的数据…
- 社团快速发展!
 - 更多的用户
 - 更多的子项目!
- 子项目 exploding
 - 给母项目带来价值
 - 发展生态系统



大量应用

	2008	2009
Webmap	<p>~70 小时用时</p> <p>~300 TB shuffling</p> <p>~200 TB 输出</p>	<p>~73 小时用时</p> <p>~490 TB shuffling</p> <p>~280 TB 输出</p> <p>+55% 硬件</p>
Sort benchmarks (Jim Gray contest)	<p>1 Terabyte 排序</p> <ul style="list-style-type: none">• 209 秒• 900 节点	<p>1 Terabyte 排序</p> <ul style="list-style-type: none">• 62 秒, 1500 节点 <p>1 Petabyte sorted</p> <ul style="list-style-type: none">• 16.25 小时, 3700 节点
Largest cluster	<p>2000 节点</p> <ul style="list-style-type: none">• 6PB 硬盘• 16TB 内存• 16K 核	<p>4000 节点</p> <ul style="list-style-type: none">• 16PB 硬盘• 64TB 内存• 32K 核• (快 40%!)