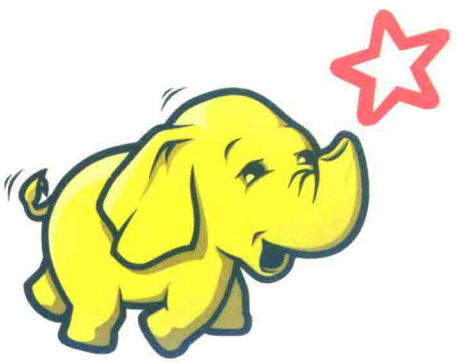
大数据技术

清华大学电子工程系教授、博士生导师 黄永峰
中国科学院大学人工智能技术学院教授、副院长 肖 俊
中国科学院计算机网络中心百人计划研究员、
巴黎第13大学客座教授、里昂第1大学客座教授 贺海武
传智播客、黑马程序员高级副总裁 方立勋
搜狐视频技术总监 杨志云
Oracle OAEC产品总监、Oracle认证高级讲师 刘 彰

联袂推荐



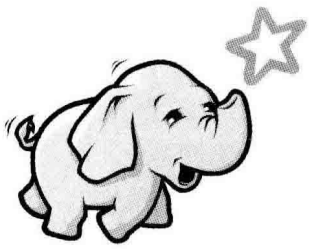
两大生态系统操作快速入门

Hadoop + Spark

生态系统操作与实战指南

余辉 著

清華大學出版社



Hadoop + Spark

生态系统操作与实战指南

余辉 著

清華大學出版社
北京

内 容 简 介

本书用于 Hadoop+Spark 快速上手，全面解析 Hadoop 和 Spark 生态系统，通过原理解说和实例操作每一个组件，让读者能够轻松跨入大数据分析与开发的大门。

全书共 12 章，大致分为 3 个部分，第 1 部分（第 1~7 章）讲解 Hadoop 的原生态组件，包括 Hadoop、ZooKeeper、HBase、Hive 环境搭建与安装，以及介绍 MapReduce、HDFS、ZooKeeper、HBase、Hive 原理和 Apache 版本环境下实战操作。第 2 部分（第 8~11 章）讲解 Spark 的原生态组件，包括 Spark Core、Spark SQL、Spark Streaming、DataFrame，以及介绍 Scala、Spark API、Spark SQL、Spark Streaming、DataFrame 原理和 CDH 版本环境下实战操作，其中 Flume 和 Kafka 属于 Apache 顶级开源项目也放在本篇讲解。第 3 部分（第 12 章）讲解两个大数据项目，包络网页日志离线项目和实时项目，在 CDH 版本环境下通过这两个项目将 Hadoop 和 Spark 原生态组件进行整合，一步步带领读者学习和实战操作。

本书适合想要快速掌握大数据技术的初学者，也适合作为高等院校和培训机构相关专业师生的教学参考书和实验用书。

图书在版编目（CIP）数据

Hadoop+Spark 生态系统操作与实战指南 / 余辉著. — 北京：清华大学出版社，2017

ISBN 978-7-302-47967-3

I. ①H… II. ①余… III. ①数据处理软件－指南 IV. ①TP274

中国版本图书馆 CIP 数据核字（2017）第 207245 号

责任编辑：夏毓彦
封面设计：王　翔
责任校对：闫秀华
责任印制：宋　林

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质 量 反 馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
印 装 者：三河市铭诚印务有限公司
经　　销：全国新华书店
开　　本：190mm×260mm　　**印　　张**：22　　**字　　数**：563 千字
版　　次：2017 年 9 月第 1 版　　**印　　次**：2017 年 9 月第1 次印刷
印　　数：1～3500
定　　价：69.00 元

产品编号：076840-01

推荐序

大数据是继石油之后，新兴的一种国家战略资源。大数据研究、开发和应用已经成为全球学术界、产业界的焦点。Hadoop、Spark 等开源项目是目前大数据领域应用最广泛的技术和平台。熟练掌握 Hadoop、Spark 等是从事大数据研发和应用等从业人员必备的基本技能。

《Hadoop＋Spark 生态系统操作与实战指南》正是在这样的技术背景下应运而生，能极好地满足广大大数据从业者的需求。本书以原理介绍为基础，以实战训练为目标，具体、深入地阐述了 Hadoop 及 Spark 的原生态中每一个组件的基本原理和应用方法；选择 Apache 和 CDH 两个主流 Hadoop 版本作为剖析实例，通过 Java、Scala、客户端等开发案例，采用主流的离线项目和实时项目进行讲解。

作者根据自己多年在大数据行业的研发经验和个人体会，并结合大数据实际研发中需求和特点，认真整理其多年来编写的有关大数据研发的博文，精心组织和修订，最终编撰此著作，馈食读者。因此，该著作既是在大数据一线研发人员的知识结晶，而且还是有意进军大数据领域的从业人员的“良师益友”，确实是一本难得的大数据研发的参考资料。

黄永峰

清华大学电子工程系教授、博士生导师

随着大数据时代的到来，大数据技术在各行各业的应用越来越多，大数据相关技术的学习和使用者也越来越多。《Hadoop+Spark 生态系统操作与实战指南》从大数据爱好者和入门者的角度出发，以原理兼实战为主体思路展现 Hadoop 及 Spark 的原生态中每一个组件的操作方法，是一本有效的快速入门教程。

本书首先讲解了 Apache 和 CDH 两大 Hadoop 版本的集群搭建，并以此作为后续的开发平台；其次，讲解了 Hadoop+Spark 中原生态组件的原理，并使用 Java、Scala、客户端对组件进行实例操作，作为案例；最后，通过两个网页日志分析项目将 Hadoop 和 Spark 中的原生态组件整合在一起，作为项目架构。

余辉毕业于中国科学院大学，其研究方向为大数据与云计算，目前已拥有多年一线大数据开发经验。本书将理论与实践相结合，可作为相关技术教学和培训的参考资料。

肖俊

中国科学院大学人工智能技术学院教授、副院长

本书系统介绍了大数据相关知识，全书共有 12 章，论述了大数据的基本概念、大数据处理架构 Hadoop、分布式文件系统 HDFS、分布式数据库 HBase、NoSQL 数据库、云数据库、分布式并行编程模型 MapReduce、基于内存的分布式计算框架 Spark、最新的 ZooKeeper、Hive、Scala、Flume、Kafka 等技术。在 Hadoop、HDFS、HBase、MapReduce 和 Spark 等重要章节，都安排了实践操作，让读者更好地学习和掌握大数据关键技术。

本文作者余辉工程师，在大数据领域的实验室及公司工作多年，积累了丰富的实战经验。这本书理论结合实践，手把手教读者一步一步入门，避免了“纸上谈兵”，是大数据研究爱好者及从业人员的入门书籍。

本书可以作为高等院校计算机专业、信息管理等相关专业的大数据课程教材，也可供相关技术人员参考、学习、培训之用。

贺海武

中国科学院计算机网络信息中心（CNIC/CAS）百人计划研究员、

巴黎第 13 大学客座教授、里昂第 1 大学客座教授

市面上有许多讲解 Hadoop 或者 Spark 的书籍，但很难找到一本能带领大数据爱好者快速入门的书籍。本书作者余辉兼职于 Oracle OAEC 在线教育集团大数据讲师，他从一个讲师的角度写书，本书通过多维度讲解 Hadoop+Spark 原生态系统组件，在平台环境方面使用到 Apache 和 CDH 版本的 Hadoop 集群，在开发环境方面使用到 Eclipse+Java 和 IntelliJ IDEA+Scala，在项目环境方面使用到主流的离线日志分析和实时日志分析，让大数据爱好者可以快速认识大数据、熟悉大数据、操作大数据、运用大数据。本书详细讲解了 Hadoop+Spark 原生态组件的原理，通过 Java、Scala、客户端等开发案例并附上图片进行解说，让读者极易上手，本书非常适合作为一本大数据的快速入门教材。

方立勋
传智播客.黑马程序员高级副总裁

我与余辉的认识起源于清华大学，当时他在清华大学电子工程系担任软件工程师一职，通过和他多次交谈感觉此人思维缜密、善于总结且非常热爱技术。此书涵盖了余辉多年的一线开发经验和博文总结。

《Hadoop + Spark 生态系统操作与实战指南》总计 12 章。涵盖 Hadoop+Spark 原生态系统组件，对每一个组件原理和架构有着清晰的描述。通过两套主流开发环境 Eclipse+Java 和 IntelliJ IDEA+Scala 以及客户端分别对每一个组件进行了大量的案例操作，并配上大量案例截图，最后采用主流的离线项目和实时项目进行生态组件的融合。从多维度让读者对大数据快速认知、快速理解、快速上手、快速深入了解大数据行业，是一本非常适合大数据开发爱好者快速入门的书籍。

杨志云
搜狐视频技术总监

大数据在各行各业的应用越来越广，近几年“大数据”一词也非常火热，余辉的书《Hadoop + Spark 生态系统操作与实战指南》生逢其时。虽然现在世面上有不少关于大数据方面的书籍，但我还是想从本书的内容结构，及我与作者交往方面，对此书及此人做一个概要性的介绍。

此书最大的特点是理念、实战与项目的结合，能把各个知识点，以实战操作的方式连成线，再以项目的方式，把各知识模块连成面，点、线、面轮廓清晰、项目实用，能帮助读者快速理解大数据生态技术中的各种技术在实际应用中的作用。Hadoop 是大数据平台，它通过一系列的技术组成一个大数据生态技术圈，各种技术在这个生态中是干什么、原理是什么等在书中都有讲解。书中内容包括三大部分，12 章，从大数据生态平台起源讲起（第 1 章），实践环境搭建（第 2 章）、分布式存储与计算框架介绍（第 3 章）、平台协作套件（第 4 章）、Apache 原生的分布式计算框架详解（第 5 章）、分布式数据存储数据库（第 6、7 章）、利用函数式编程处理数据（第 8 章）、数据同步（第 9、10 章）、内存计算引擎架构（第 11 章）以及综合项目（第 12 章），内容丰富、案例真实、可操作性强，通过本书，读者能快速地理解 Hadoop 大数据技术生态中各种技术在实际项目中的应用。

关于此人，余辉是我通过 CSDN 博客找到他的，最开始我是阅读他的 CSDN 博文，从他的博文字里行间能感受到他的几种特质：专注、坚持、超强的执行力。因为 Oracle OAEC 人才产业中心此时正在开设大数据相关的课程，所以通过电话联系到他，经过一段时间的交流，最终成为 Oracle OAEC 人才产业基地的一名大数据兼职讲师，负责北京 Oracle OAEC 中心的大数据课程的教授。在教学过程中，得到学员的多次好评，以此基础，我建议他写一本关于这方面的书籍。我的逻辑是让他通过授课的方式，将多年在大数据一线的实际应用与项目，用通俗易懂的方式让学员理解；同时，自己也加深了理解；再通过写书的方式，能系统地将知识、经验、和自己的理解分享给别人。

刘 彰
Oracle OAEC 人才产业集团大数据学院与认证中心产品总监
ORACLE 认证高级讲师

前　言

近几年来，随着计算机和信息技术的迅猛发展和普及应用，行业应用系统的规模迅速扩大，行业应用所产生的数据呈爆炸性增长。大数据技术快速火热，大数据开发工程师更是供不应求。本书是一本 Hadoop+Spark 快速上手的书，从 Hadoop 生态系统和 Spark 生态系统全面原理解析和实战操作每一个组件，每一个知识点都讲得十分细致，让读者能够轻松地跨入大数据开发工程师的大门。

大数据工程师薪资

近几年大数据岗位尤其火热，大数据开发工程师供不应求，市面上大数据开发工程师起步就是 8 千元，1 年工作经验 1 万 2 千元，2 年工作经验 1 万 5 千元，3 年工作经验 2 万以上。根据每个人自身学习能力不同，有人 2 年就可以达到 2 万元以上。

下图是神州数码于 2017 年 6 月 6 日发布的一则招聘信息。

神州数码招聘

大数据开发工程师

15k-20k /北京 / 经验1-3年 / 本科及以上 / 全职

大数据　数据库　大数据

2天前 发布于拉勾网

职位诱惑：

福利优厚 扁平化管理 正编岗 环境舒适

职位描述：

岗位说明：

1、负责电商平台数据相关的开发工作；

2、预研大数据领域的前沿技术，优化现有架构。

岗位要求：

1、计算机及相关专业本科以上学历；

2、熟悉Java、Jvm，熟悉网络编程、多线程等；

3、有Linux操作系统使用、开发经验，熟悉hadoop相关技术框架，比如zookeeper、hive、hbase、spark等分布式运算存储框架；

4、熟练使用MySql、Hbase关系型及非关系型数据库的操作；

5、参与过大数据平台或分布式系统开发工作；

6、优秀的学习能力，扎实的数据结构和算法基础；

7、优秀的分析、解决问题的能力，对挑战性问题充满激情；

8、具备强烈的工作责任感，具备良好的团队精神和沟通能力。

本书内容

全书共 12 章，分为 3 个部分，第 1 部分（第 1~7 章）讲解了 Hadoop 的原生态组件，包括 Hadoop、ZooKeeper、HBase、Hive 环境搭建与安装，以及如何对 MapReduce、HDFS、ZooKeeper、HBase、Hive 进行原理介绍和 Apache 版本环境下实战的操作。第 2 部分（第 8~11 章）讲解 Spark 的原生态组件，包括 Spark Core、Spark SQL、Spark Streaming、DataFrame，以及如何对 Scala、Spark API、Spark SQL、Spark Streaming、DataFrame 进行原理介绍和 CDH 版本环境下实战的操作，其中 Flume 和 Kafka 属于 Apache 顶级开源项目也放在本篇讲解。第 3 部分（第 12 章）讲解大数据项目，包络网页日志离线项目和实时项目，在 CDH 版本环境下通过两个项目将 Hadoop 和 Spark 原生态组件进行整合，一步步带领读者实战大数据项目。

本书特色

本书是一本 Hadoop + Spark 的快速入门书籍，以通俗易懂的方式介绍了 Hadoop + Spark 原生态组件的原理、实战操作以及集群搭建方面的知识。其中，Hadoop 原生态组件包括：MapReduce、HDFS、ZooKeeper、HBase、Hive；Spark 原生态组件包括：Spark Core、Spark SQL、Spark Streaming、Dataframe；同时包括 Apache 版本和 CDH5 版本的 Hadoop 集群搭建。本书的特点是：注重“实战”训练，强调知识系统性，关注内容实用性。

（1）本书从培训角度对读者简述 Hadoop + Spark 中常用组件的原理和实战操作，让读者快速了解组件原理和功能使用。

（2）每一个操作都配有实例代码或者图片来帮助理解，每一章的最后还有小节，以归纳总结本章的内容，帮助读者对 Hadoop + Spark 原生态系统有一个大的全局观。

（3）目前市面上关于 Hadoop 的书很多，关于 Spark 的书也很多，但是很少有对 Hadoop + Spark 结合进行讲解。本书首先讲解 Hadoop + Spark 原理，接着讲解 Hadoop + Spark 原生态组件的实例操作，最后结合大数据网站日志离线和实时两个项目融合 Hadoop+Spark 所有生态系统功能，使读者对本书有一个由浅入深且快速上手的过程。

本书适合读者

本书适合 Hadoop+Spark 的初学者，希望深入了解 Hadoop+Spark 安装部署、开发优化的大数据工程师，希望深入了解 Hadoop+Spark 管理、业务框架扩展的大数据架构师，以及任何对 Hadoop+Spark 相关技术感兴趣的读者。

本书代码、软件、文档下载

本书代码、软件、文档下载地址（注意数字和字母大小写）如下：

http://pan.baidu.com/s/1cCi0k2

如果下载有问题，请联系电子邮箱 booksaga@163.com，邮件主题为“Hadoop+Spark 生态系统与实战指南”。

本书作者

余辉，中国科学院大学硕士研究生毕业，研究方向为云计算和大数据。现供职于某上市公司技术经理，并在 Oracle OAEC 人才产业集团大数据学院（http://www.oracleoaec.com.cn/）担任大数据讲师。曾在清华大学电子工程系 NGNLab 研究室（http://ngn.ee.tsinghua.edu.cn/）担任软件工程师。

已发表两篇大数据论文：《微博舆情的 Hadoop 存储和管理平台设计与实现》和《跨媒体多源网络舆情分析系统设计与实现》

博客：http://blog.csdn.net/silentwolfyh

微博：http://weibo.com/u/3195228233

电子邮箱：yuhuiqh2009@163.com

致谢

赶在儿子 1 岁生日之际，赶在我告别 30 岁之际，我撰写《Hadoop+Spark 生态系统操作与实战指南》一书，作为我儿子的生日礼物。感谢父母提供了良好的生活环境，感谢舅舅、舅妈提供了良好的学习平台，感谢我的老婆、姐姐、姐夫在生活上对我的支持和奉献。最后，感谢清华工作和学习的那些时光，清华六年，我学会了生存技能、找到了研究方向、培养了生活习惯。

余 辉

2017 年 7 月

目 录

第 1 章

◂ Hadoop概述 ▸

1.1 Hadoop 简介

1. Hadoop 的由来

Hadoop 是 Doug Cutting（Apache Lucene 创始人）开发的、使用广泛的文本搜索库。Hadoop 起源于 Apache Nutch，后者是一个开源的网络搜索引擎，本身也是 Lucene 项目的一部分。

2. Hadoop 名字的起源

Hadoop 这个名字不是一个缩写，它是一个虚构的名字。该项目的创建者 Doug Cutting 如此解释 Hadoop 的得名："这个名字是我孩子给一头吃饱了的棕黄色大象命名的。我的命名标准就是简短、容易发音和拼写，没有太多的意义，并且不会被用于别处。小孩子是这方面的高手。Googol 就是由小孩命名的。"（Google 来源于 Googol 一词。GooGol 指的是 10 的 100 次幂（方），代表互联网上的海量资源。公司创建之初，肖恩·安德森在搜索该名字是否已经被注册时，将 Googol 误打成了 Google。）

Hadoop 及其子项目和后继模块所使用的名字往往也与其功能不相关，经常用一头大象或其他动物主题（例如：Pig）。较小的各个组成部分给予更多描述性（因此也更俗）的名称。这是一个很好的原则，因为它意味着可以大致从其名字猜测其功能，例如，jobtracker 的任务就是跟踪 MapReduce 作业。

从头开始构建一个网络搜索引擎是一个雄心勃勃的目标，不只是要编写一个复杂的、能够抓取和索引网站的软件，还需要面临着没有专业运行团队支持运行它的挑战，因为它有那么多独立部件。同样昂贵的还有：据 Mike Cafarella 和 Doug Cutting 估计，一个支持此 10 亿页的索引，需要价值约 50 万美元的硬件投入，每月运行费用还需要 3 万美元。 不过，他们相信这是一个有价值的目标，因为这会开放并最终使搜索引擎算法普及化。

Nutch 项目开始于 2002 年，一个可工作的抓取工具和搜索系统很快浮出水面。但他们意识到，他们的架构将无法扩展到拥有数十亿网页的网络。在 2003 年发表的一篇描述 Google 分布式文件系统（简称 GFS）的论文为他们提供了及时的帮助，文中称 Google 正在使用此文件系统。GFS 或类似的东西，可以解决他们在网络抓取和索引过程中产生的大量的文件的存储需求。具体而言，GFS 会省掉管理所花的时间，如管理存储节点。在 2004 年，他们开

始写一个开放源码的应用，即 Nutch 的分布式文件系统（NDFS）。

2004 年，Google 发表了论文，向全世界介绍了 MapReduce。2005 年初，Nutch 的开发者在 Nutch 上有了一个可工作的 MapReduce 应用，到当年年中，所有主要的 Nutch 算法被移植到使用 MapReduce 和 NDFS 来运行。

Nutch 中的 NDFS 和 MapReduce 实现的应用远不只是搜索领域，在 2006 年 2 月，他们从 Nutch 转移出来成为一个独立的 Lucene 子项目，称为 Hadoop。大约在同一时间，Doug Cutting 加入雅虎，Yahoo 提供一个专门的团队和资源将 Hadoop 发展成一个可在网络上运行的系统（见后文的补充材料）。在 2008 年 2 月，雅虎宣布其搜索引擎产品部署在一个拥有 1 万个内核的 Hadoop 集群上。

2008 年 1 月，Hadoop 已成为 Apache 顶级项目，证明它是成功的，是一个多样化、活跃的社区。通过这次机会，Hadoop 成功地被雅虎之外的很多公司应用，如 Last.fm、Facebook 和《纽约时报》。一些应用在 Hadoop 维基有介绍，Hadoop 维基的网址为 http://wiki.apache.org/hadoop/PoweredBy。

有一个良好的宣传范例，《纽约时报》使用亚马逊的 EC2 云计算将 4 TB 的报纸扫描文档压缩，转换为用于 Web 的 PDF 文件。这个过程历时不到 24 小时，使用 100 台机器运行，如果不结合亚马逊的按小时付费的模式（即允许《纽约时报》在很短的一段时间内访问大量机器）和 Hadoop 易于使用的并行程序设计模型，该项目很可能不会这么快开始启动。

2008 年 4 月，Hadoop 打破世界纪录，成为最快排序 1 TB 数据的系统，运行在一个 910 节点的集群，Hadoop 在 209 秒内排序了 1 TB 的数据（还不到三分半钟），击败了前一年的 297 秒冠军。同年 11 月，谷歌在报告中声称，它的 MapReduce 实现执行 1 TB 数据的排序只用了 68 秒。在 2009 年 5 月，有报道宣称 Yahoo 的团队使用 Hadoop 对 1 TB 的数据进行排序只花了 62 秒时间。

构建互联网规模的搜索引擎需要大量的数据，因此需要大量的机器来进行处理。Yahoo！Search 包括四个主要组成部分：Crawler，从因特网下载网页；WebMap，构建一个网络地图；Indexer，为最佳页面构建一个反向索引；Runtime（运行时），回答用户的查询。WebMap 是一幅图，大约包括一万亿条边（每条代表一个网络链接）和一千亿个节点（每个节点代表不同的网址）。创建和分析此类大图需要大量计算机运行若干天。在 2005 年初，WebMap 所用的基础设施名为 Dreadnaught，需要重新设计以适应更多节点的需求。Dreadnaught 成功地从 20 个节点扩展到 600 个，但还需要一个完全重新的设计，以进一步扩大。Dreadnaught 与 MapReduce 有许多相似的地方，但灵活性更强，结构更少。具体说来，Dreadnaught 作业可以将输出发送到此作业下一阶段中的每一个分段（fragment），但排序是在库函数中完成的。在实际情形中，大多数 WebMap 阶段都是成对存在的，对应于 MapReduce。因此，WebMap 应用并不需要为了适应 MapReduce 而进行大量重构。

Eric Baldeschwieler（Eric14）组建了一个小团队，他们开始设计并原型化一个新的框架（原型为 GFS 和 MapReduce，用 C++语言编写），打算用它来替换 Dreadnaught。尽管当务之急是需要一个 WebMap 新框架，但显然，标准化对于整个 Yahoo! Search 平台至关重要，并且通过使这个框架泛化，足以支持其他用户，这样他们才能够充分运用对整个平台的投资。

与此同时，雅虎在关注 Hadoop（当时还是 Nutch 的一部分）及其进展情况。2006 年 1

月，雅虎聘请了 Doug Cutting，一个月后，决定放弃自己的原型，转而使用 Hadoop。相较于雅虎自己的原型和设计，Hadoop 的优势在于它已经在 20 个节点上实际应用过。这样一来，雅虎便能在两个月内搭建一个研究集群，并着手帮助真正的客户使用这个新的框架，速度比原来预计的快许多。另一个明显的优点是 Hadoop 已经开源，较容易（虽然远没有那么容易！）从雅虎法务部门获得许可在开源方面进行工作。因此，雅虎在 2006 年初设立了一个 200 个节点的研究集群，他们将 WebMap 的计划暂时搁置，转而为研究用户支持和发展 Hadoop。

3. Hadoop 大事记

2004 年，最初的版本（现在称为 HDFS 和 MapReduce）由 Doug Cutting 和 Mike Cafarella 开始实施。

2005 年 12 月，Nutch 移植到新的框架，Hadoop 在 20 个节点上稳定运行。

2006 年 1 月，Doug Cutting 加入雅虎。

2006 年 2 月，Apache Hadoop 项目正式启动以支持 MapReduce 和 HDFS 的独立发展。

2006 年 2 月，雅虎的网格计算团队采用 Hadoop。

2006 年 4 月，标准排序（10 GB 每个节点）在 188 个节点上运行 47.9 个小时。

2006 年 5 月，雅虎建立了一个 300 个节点的 Hadoop 研究集群。

2006 年 5 月，标准排序在 500 个节点上运行 42 个小时（硬件配置比 4 月的更好）。

2006 年 11 月，研究集群增加到 600 个节点。

2006 年 12 月，标准排序在 20 个节点上运行 1.8 个小时，100 个节点 3.3 小时，500 个节点 5.2 小时，900 个节点 7.8 个小时。

2007 年 1 月，研究集群到达 900 个节点。

2007 年 4 月，研究集群达到两个 1000 个节点的集群。

2008 年 4 月，赢得世界最快 1 TB 数据排序在 900 个节点上用时 209 秒。

2008 年 10 月，研究集群每天装载 10 TB 的数据。

2009 年 3 月，17 个集群总共 24 000 台机器。

2009 年 4 月，赢得每分钟排序，59 秒内排序 500 GB（在 1400 个节点上）和 173 分钟内排序 100 TB 数据（在 3400 个节点上）。

1.2 Hadoop 版本和生态系统

1. Hadoop 版本的优缺点

目前市面上 Hadoop 版本主要有两种：Apache 版本和 CDH 版本。

（1）Aapche 版本的 Hadoop

官网：http://hadoop.apache.org/

Aapche Hadoop 优势：对硬件要求低。

Aapche Hadoop 劣势：搭建烦琐，维护烦琐，升级烦琐，添加组件烦琐。