



大数据知识体系型公开课 《大数据概念、技术与应用》

第2讲 大数据处理架构Hadoop

林子雨 博士/助理教授

厦门大学计算机科学系

厦门大学云计算与大数据研究中心

E-mail: ziyulin@xmu.edu.cn ▶▶

主页: <http://www.cs.xmu.edu.cn/linziyu>

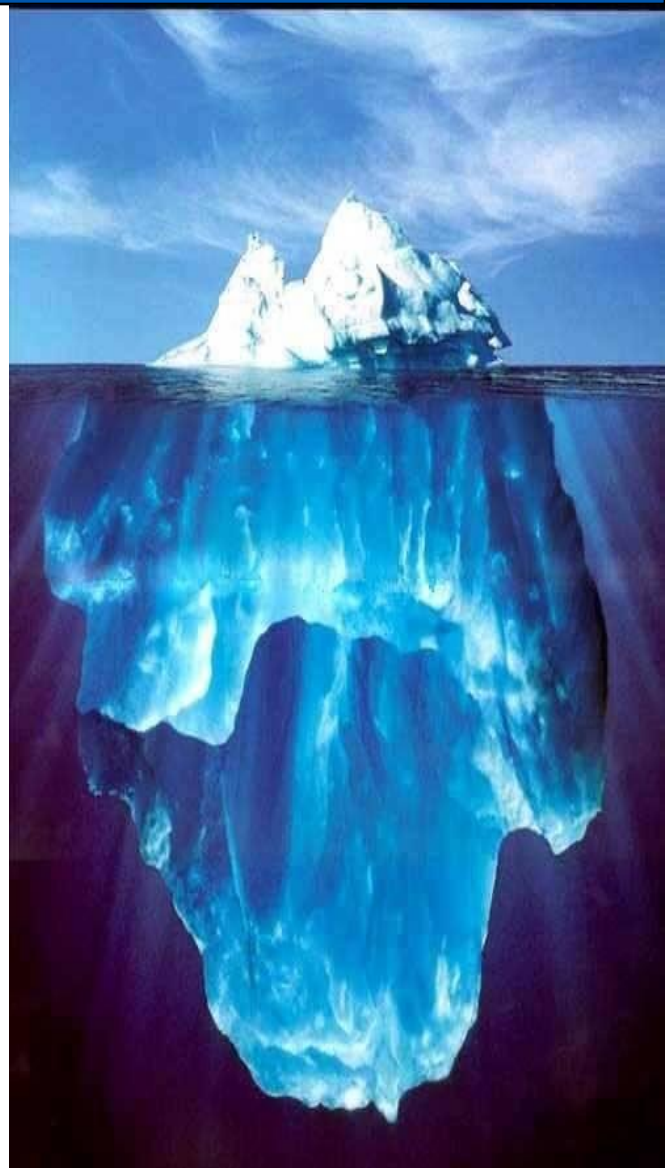




提纲

- 2.1 概述
- 2.2 Hadoop项目结构

欢迎访问《大数据技术原理与应用》教材官方网站：
<http://dmlab.xmu.edu.cn/post/bigdata>





2.1 概述

- 2.1.1 Hadoop简介
- 2.1.2 Hadoop发展简史
- 2.1.3 Hadoop的特性
- 2.1.4 Hadoop的应用现状



2.1.1 Hadoop简介

- Hadoop是Apache软件基金会旗下的一个开源分布式计算平台，为用户提供了系统底层细节透明的分布式基础架构
- Hadoop是基于Java语言开发的，具有很好的跨平台特性，并且可以部署在廉价的计算机集群中
- Hadoop的核心是分布式文件系统HDFS（Hadoop Distributed File System）和MapReduce
- Hadoop被公认为行业大数据标准开源软件，在分布式环境下提供了海量数据的处理能力
- 几乎所有主流厂商都围绕Hadoop提供开发工具、开源软件、商业化工具和技术服务，如谷歌、雅虎、微软、思科、淘宝等，都支持Hadoop



2.1.2 Hadoop发展简史



Hadoop的标志

- Hadoop最初是由Apache Lucene项目的创始人Doug Cutting开发的文本搜索库。Hadoop源自始于2002年的Apache Nutch项目——一个开源的网络搜索引擎并且也是Lucene项目的一部分
- 在2004年，Nutch项目也模仿GFS开发了自己的分布式文件系统NDFS（Nutch Distributed File System），也就是HDFS的前身
- 2004年，谷歌公司又发表了另一篇具有深远影响的论文，阐述了MapReduce分布式编程思想
- 2005年，Nutch开源实现了谷歌的MapReduce



2.1.2 Hadoop发展简史

- 到了2006年2月，Nutch中的NDFS和MapReduce开始独立出来，成为Lucene项目的一个子项目，称为Hadoop，同时，Doug Cutting加盟雅虎
- 2008年1月，Hadoop正式成为Apache顶级项目，Hadoop也逐渐开始被雅虎之外的其他公司使用
- 2008年4月，Hadoop打破世界纪录，成为最快排序1TB数据的系统，它采用一个由910个节点构成的集群进行运算，排序时间只用了209秒
- 在2009年5月，Hadoop更是把1TB数据排序时间缩短到62秒。Hadoop从此名声大震，迅速发展成为大数据时代最具影响力的开源分布式开发平台，并成为事实上的大数据处理标准



2.1.3 Hadoop的特性

Hadoop是一个能够对大量数据进行分布式处理的软件框架，并且是以一种可靠、高效、可伸缩的方式进行处理，它具有以下几个方面的特性：

- 高可靠性
- 高效性
- 高可扩展性
- 高容错性
- 成本低
- 运行在Linux平台上
- 支持多种编程语言



2.1.3 Hadoop的应用现状

- Hadoop凭借其突出的优势，已经在各个领域得到了广泛的应用，而互联网领域是其应用的主阵地
- 2007年，雅虎在Sunnyvale总部建立了M45——一个包含了4000个处理器和1.5PB容量的Hadoop集群系统
- Facebook作为全球知名的社交网站，Hadoop是非常理想的选择，Facebook主要将Hadoop平台用于日志处理、推荐系统和数据仓库等方面
- 国内采用Hadoop的公司主要有百度、淘宝、网易、华为、中国移动等，其中，淘宝的Hadoop集群比较大



2.2 Hadoop项目结构

- 经过多年的发展，Hadoop项目不断完善和成熟，目前已经包含多个子项目（如图2-2所示）
- 除了核心的HDFS和MapReduce以外，Hadoop项目还包括Common、Avro、Zookeeper、HBase、Hive、Chukwa、Pig等子项目，它们提供了互补性服务或在核心层上提供了更高层的服务

Pig	Chukwa	Hive	HBase
MapReduce	HDFS	ZooKeeper	
Common	Avro		

图2-2 Hadoop项目结构图



本讲小结

- Hadoop被视为事实上的大数据处理标准，本章介绍了Hadoop的发展历程，并阐述了Hadoop的高可靠性、高效性、高可扩展性、高容错性、成本低、运行在Linux平台上、支持多种编程语言等特性
- Hadoop目前已经在各个领域得到了广泛的应用，雅虎、Facebook、百度、淘宝、网易等公司都建立了自己的Hadoop集群
- 经过多年发展，Hadoop项目已经变得非常成熟和完善，包括Common、Avro、Zookeeper、HDFS、MapReduce、HBase、Hive、Chukwa、Pig等子项目，其中，HDFS和MapReduce是Hadoop的两大核心组件
- 本章最后介绍了如何在Linux系统下完成Hadoop的安装和配置，这个部分是后续章节实践环节的基础



主讲教师



主讲教师：林子雨

单位：厦门大学计算机科学系

E-mail: ziyulin@xmu.edu.cn

个人网页: <http://www.cs.xmu.edu.cn/linziyu>

数据库实验室网站: <http://dblab.xmu.edu.cn>



扫一扫访问个人主页

林子雨，男，1978年出生，博士（毕业于北京大学），现为厦门大学计算机科学系助理教授（讲师），曾任厦门大学信息科学与技术学院院长助理、晋江市发展和改革委员会副局长。中国高校首个“数字教师”提出者和建设者，厦门大学数据库实验室负责人，厦门大学云计算与大数据研究中心主要建设者和骨干成员，2013年度厦门大学奖教金获得者。主要研究方向为数据库、数据仓库、数据挖掘、大数据、云计算和物联网，编著出版中国高校第一本系统介绍大数据知识的专业教材《大数据技术原理与应用》并成为畅销书籍；主讲厦门大学计算机系本科生课程《数据库系统原理》和研究生课程《分布式数据库》《大数据技术基础》。具有丰富的政府和企业信息化培训经验，曾先后给中国移动通信集团公司、福州马尾区政府、福建省物联网科学研究院、石狮市物流协会、厦门市物流协会等多家单位和个人开展信息化培训，累计培训人数达2000人以上。



大数据学习教材推荐



扫一扫访问教材官网

《大数据技术原理与应用——概念、存储、处理、分析与应用》，由厦门大学计算机科学系林子雨博士编著，是中国高校第一本系统介绍大数据知识的专业教材。

全书共有13章，系统地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、流计算、图计算、数据可视化以及大数据在互联网、生物医学和物流等各个领域的应用。在Hadoop、HDFS、HBase和MapReduce等重要章节，安排了入门级的实践操作，让读者更好地学习和掌握大数据关键技术。

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。

欢迎访问《大数据技术原理与应用——概念、存储、处理、分析与应用》教材官方网站：
<http://dblab.xmu.edu.cn/post/bigdata>



Principles and Applications of Big Data Technology - Big Data Conception, Storage, Processing, Analysis and Application

林子雨 编著



- 搭建起通向“大数据知识空间”的桥梁和纽带
- 构建知识体系、阐明基本原理、引导初级实践、了解相关应用
- 为读者在大数据领域“深耕细作”奠定基础、指明方向

中国工信出版集团

人民邮电出版社
POSTS & TELECOM PRESS

The background of the slide features several faint, light-blue silhouettes of people. At the top, there are two groups of people standing and holding hands. On the right side, a person is shown in profile, looking towards the center. On the left side, two people are shown in profile, facing each other. The overall scene suggests a group of people in a meeting or a presentation.

Thank You!

Department of Computer Science, Xiamen University