厦门大学计算机科学系研究生课程

《大数据技术基础》

第3 章 Hadoop

（2013年新版）

林子雨

厦门大学计算机科学系

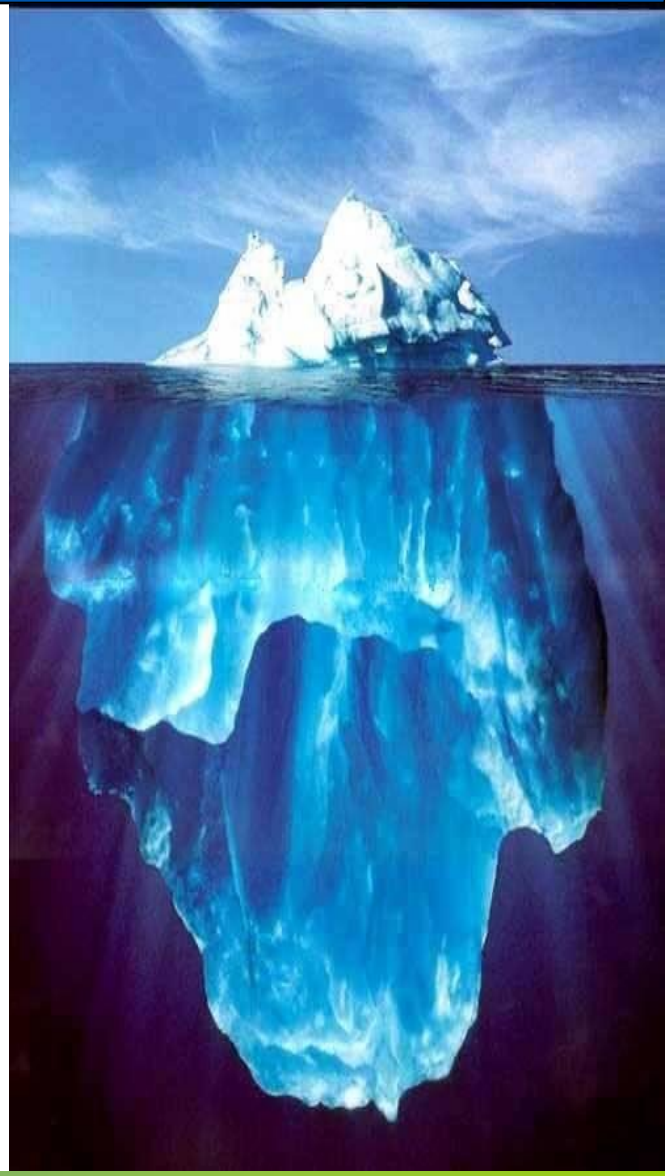
E-mail: ziyulin@xmu.edu.cn

主页：http://www.cs.xmu.edu.cn/linziyu

提纲

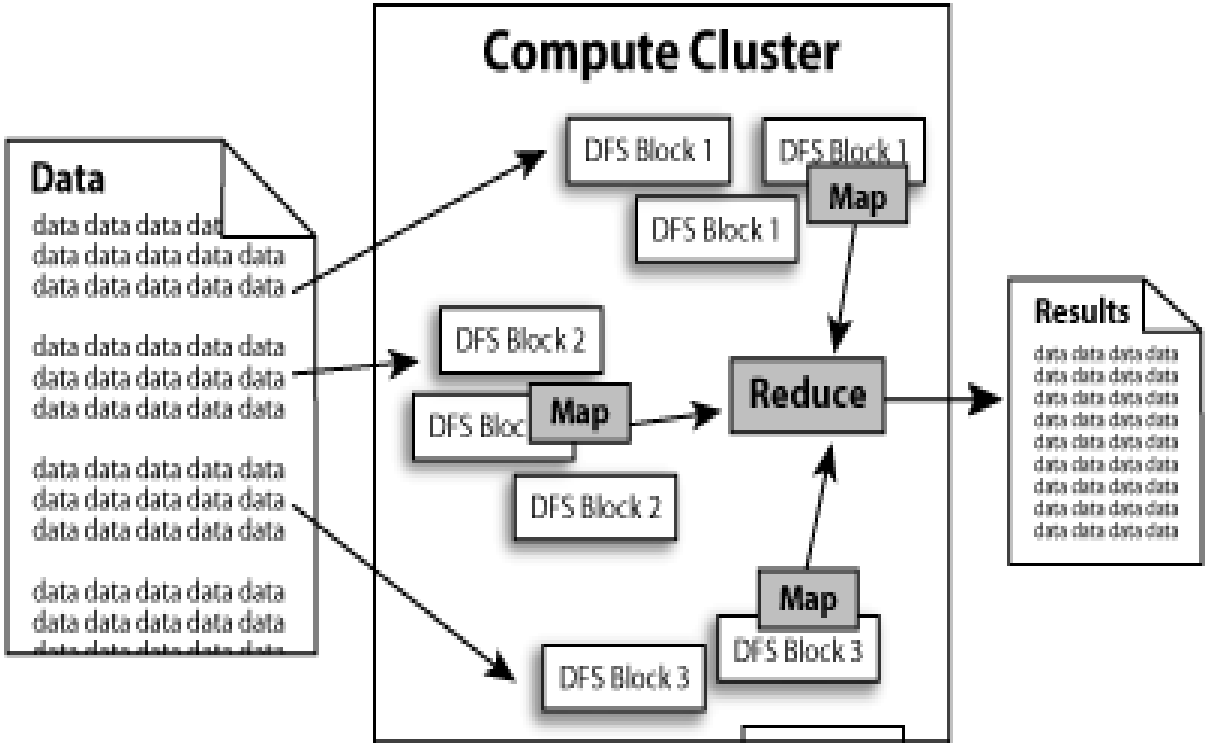
- **Hadoop概述**
- **Hadoop发展简史**
- **Hadoop的功能与作用**
- **为什么不用关系型数据库管理系统**
- **Hadoop优点**
- **Hadoop的应用现状和发展趋势**
- **Hadoop项目及其结构**
- **Hadoop的体系结构**
- **Hadoop与分布式开发**
- **Hadoop应用案例**
- **Hadoop平台上的海量数据排序**

本讲义PPT存在配套教材，由林子雨通过大量阅读、收集、整理各种资料后编写而成
下载配套教材请访问《大数据技术基础》2013班级网站：http://dblab.xmu.edu.cn/node/423



Hadoop概述

Hadoop是一个开源的可运行于大规模集群上的分布式并行编程框架，它实现了 Map/Reduce 计算模型。借助于 Hadoop， 程序员可以轻松地编写分布式并行程序，将其运行于计算机集群上，完成海量数据的计算。



Hadoop发展简史

- Hadoop起源于Apache Nutch，后者是一个开源的网络搜索引擎，本身也是由Lucene项目的一部分。
- 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。
- 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，成为Hadoop。
- 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4月，Hadoop打破世界纪录，成为最快排序1TB数据的系统。运行在一个910节点的群集，Hadoop在209秒内排序了1TB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，谷歌在报告中生成，它的MapReduce实现执行1TB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1TB的数据进行排序只花了62秒时间。

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——雅虎的网络计算团队采用Hadoop。

2006年4月——标准排序（10GB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。

Hadoop大事记

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1TB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10TB的数据。

2009年3月——17个集群总共24000台机器。

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

Hadoop的作用与功能

- Hadoop采用了分布式存储方式，提高了读写速度，并扩大了存储容量。采用MapReduce 来整合分布式文件系统上的数据，可保证分析和处理数据的高效。与此同时，Hadoop 还采用存储冗余数据的方式保证了数据的安全性。

- Hadoop中HDFS 的高容错特性，以及它是基于Java 语言开发的，这使得Hadoop可以部署在低廉的计算机集群中，同时不限于某个操作系统。Hadoop 中HDFS 的数据管理能力，MapReduce 处理任务时的高效率，以及它的开源特性，使其在同类的分布式系统中大放异彩，并在众多行业和科研领域中被广泛采用。

为什么不用关系型数据库管理系统

- 在更新一小部分数据库记录的时候，传统RDBMS采用的B树效果很好。但在更新大部分数据库数据的时候，B树的效率就没有MapReduce的效率高，因为它需要使用排序/合并来重建数据库。
- RDBMS适合点查询和更新，MapReduce适合批处理
- RDBMS适合持续更新的数据集， MapReduce适合数据被一次写入多次读取的应用
- RDBMS只能处理结构化数据，MapReduce对于非结构化或半结构化数据非常有效（避免规范化带来的非本地读问题）。
- 二者互相融合是一种趋势

	传统关系型数据库	MapReduce
数据大小	GB	PB
访问	交互型和批处理	批处理
更新	多次读写	一次写入多次读取
结构	静态模式	动态模式
集成度	高	低
伸缩性	非线性	线性

Hadoop的优点

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的，具有以下优点：

Hadoop 是可靠的：因为它假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理。

Hadoop 是高效的：因为它以并行的方式工作，通过并行处理加快处理速度。Hadoop 还是可伸缩的，能够处理 PB 级数据。

Hadoop成本低：依赖于廉价服务器：因此它的成本比较低，任何人都可以使用。

运行在Linux平台上：Hadoop带有用 Java 语言编写的框架，因此运行在 Linux 生产平台上是非常理想的。

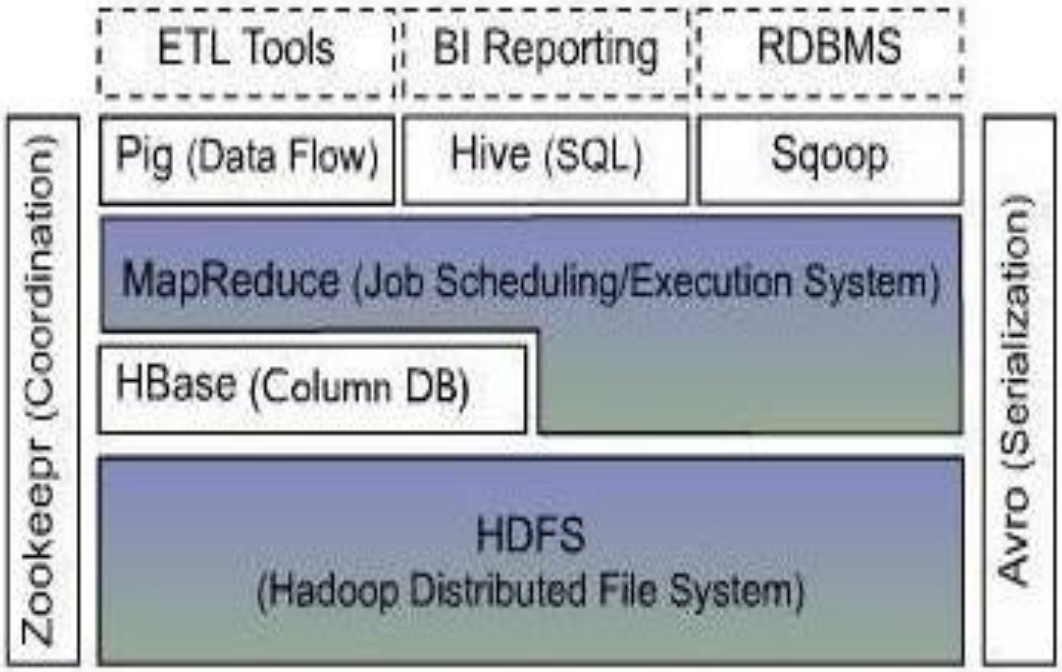
支持多种编程语言：Hadoop 上的应用程序也可以使用其他语言编写，比如 C++。

Hadoop的应用现状和发展趋势

- 由于 Hadoop 优势突出，基于Hadoop 的应用已经遍地开花，尤其是在互联网领域。
  - –Yahoo! 通过集群运行Hadoop，以支持广告系统和Web搜索的研究；
  - –Facebook 借助集群运行Hadoop，以支持其数据分析和机器学习；
  - –百度则使用Hadoop 进行搜索日志的分析和网页数据的挖掘工作；
  - –淘宝的Hadoop 系统用于存储并处理电子商务交易的相关数据；
  - –中国移动研究院基于Hadoop 的“大云”（BigCloud）系统用于对数据进行分析和并对外提供服务。
- Hadoop 目前已经取得了非常突出的成绩。随着互联网的发展，新的业务模式还将不断涌现，Hadoop 的应用也会从互联网领域向电信、电子商务、银行、生物制药等领域拓展。相信在未来，Hadoop 将会在更多的领域中扮演幕后英雄，为我们提供更加快捷优质的服务。

Hadoop项目及其结构

Hadoop有许多元素构成。最底部是 Hadoop分布式文件系统（HDFS），它存储 Hadoop集群中所有存储节点上的文件。HDFS的上一层是MapReduce引擎，该引擎由 JobTrackers 和 TaskTrackers 组成。下图描述了Hadoop生态系统中的各层子系统。

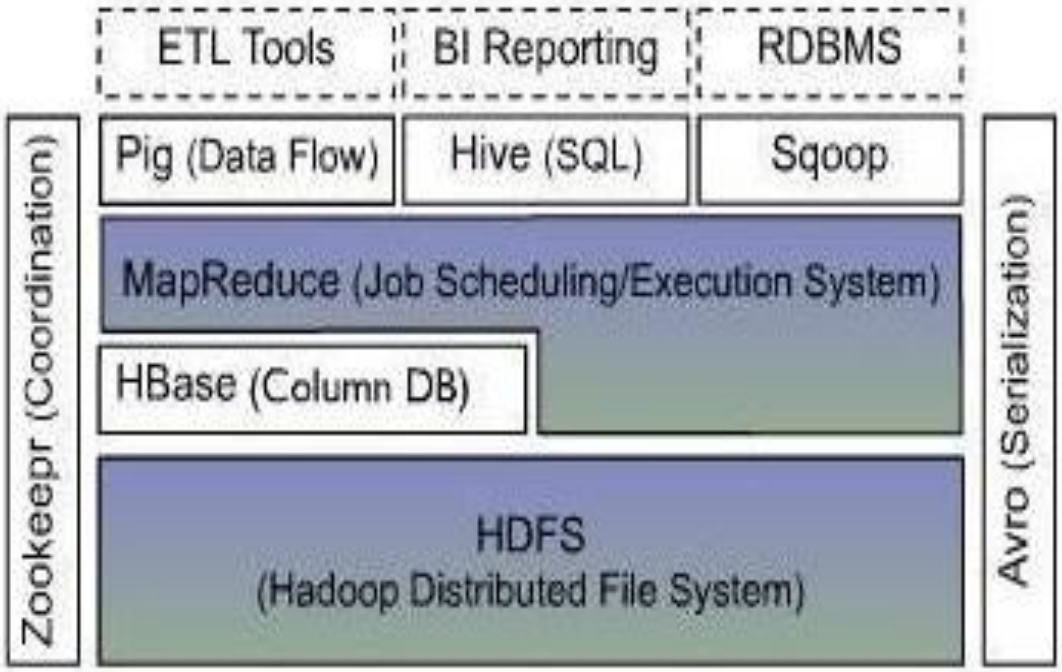


Hadoop项目及其结构

- Avro用于数据序列化的系统；
- HDFS是一种分布式文件系统，运行于大型商用机集群，HDFS为HBase提供了高可靠性的底层存储支持；
- HBase位于结构化存储层，是一个分布式的列存储数据库；
- MapReduce是一种分布式数据处理模式和执行环境，为HBase提供了高性能的计算能力；
- Zookeeper是一个分布式的、高可用性的协调服务，提供分布式锁之类的基本服务，用于构建分布式应用，为HBase提供了稳定服务和failover机制；
- Hive是一个建立在Hadoop 基础之上的数据仓库，它提供了一些用于数据整理、特殊查询和分析存储在Hadoop 文件中的数据集的工具；
- Pig是一种数据流语言和运行环境，用以检索非常大的数据集，大大简化了Hadoop常见的工作任务；
- Sqoop为HBase提供了方便的RDBMS数据导入功能，使得传统数据库数据向HBase中迁移变的非常方便。

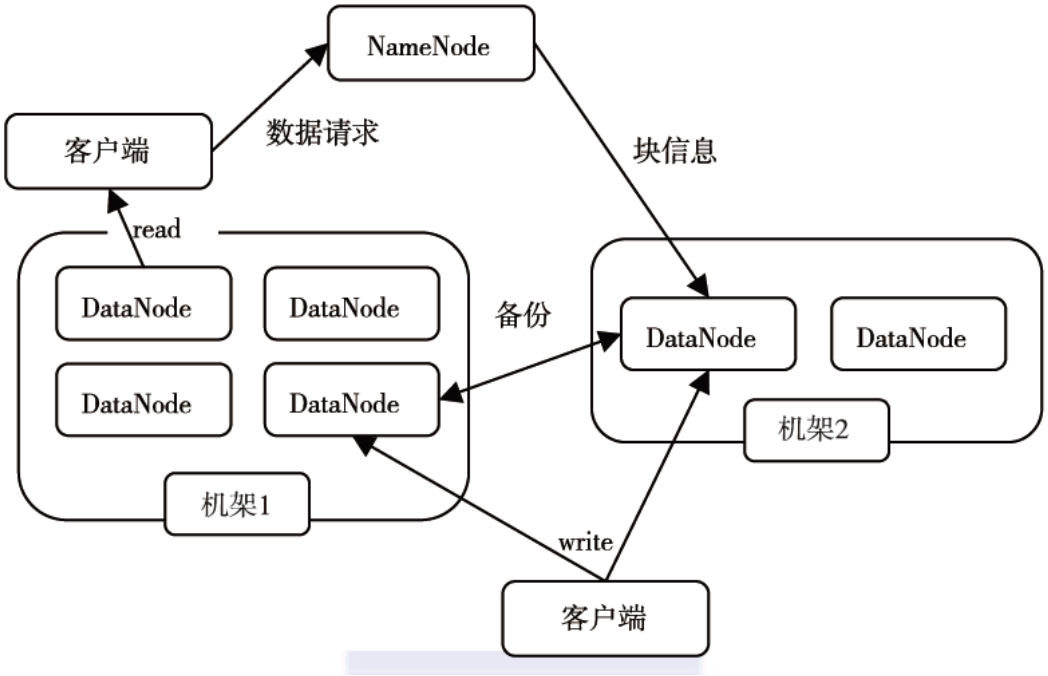
Hadoop的体系结构

HDFS和MapReduce是Hadoop的两大核心。而整个Hadoop的体系结构主要是通过HDFS 来实现对分布式存储的底层支持的，并且它会通过MapReduce 来实现对分布式并行任务处理的程序支持。



HDFS的体系结构

一个HDFS集群是由一个NameNode和若干个DataNode组成的。其中NameNode作为主服务器，管理文件系统的命名空间和客户端对文件的访问操作；集群中的DataNode管理存储的数据。HDFS 允许用户以文件的形式存储数据。从内部来看，文件被分成若干个数据块，而且这若干个数据块存放在一组DataNode上。下图给出了HDFS 的体系结构。



MapReduce的体系结构

MapReduce是一种并行编程模式，这种模式使得软件开发者可以轻松地编写出分布式并行程序。

在Hadoop的体系结构中，MapReduce是一个简单易用的软件框架，基于它可以将任务分发到由上千台商用机器组成的集群上，并以一种高容错的方式并行处理大量的数据集，实现Hadoop的并行任务处理功能。

MapReduce框架是由一个单独运行在主节点上的JobTracker 和运行在每个集群从节点上的TaskTracker共同组成的。主节点负责调度构成一个作业的所有任务，这些任务分布在不同的从节点上。主节点监控它们的执行情况，并且重新执行之前失败的任务；从节点仅负责由主节点指派的任务。当一个Job 被提交时，JobTracker接收到提交作业和配置信息之后，就会将配置信息等分发给从节点，同时调度任务并监控TaskTracker的执行。

Hadoop的体系结构

从上面的介绍可以看出，HDFS和MapReduce共同组成了Hadoop分布式系统体系结构的核心。HDFS在集群上实现了分布式文件系统，MapReduce 在集群上实现了分布式计算和任务处理。HDFS在MapReduce任务处理过程中提供了文件操作和存储等支持，MapReduce在HDFS的基础上实现了任务的分发、跟踪、执行等工作，并收集结果，二者相互作用，完成了Hadoop分布式集群的主要任务。

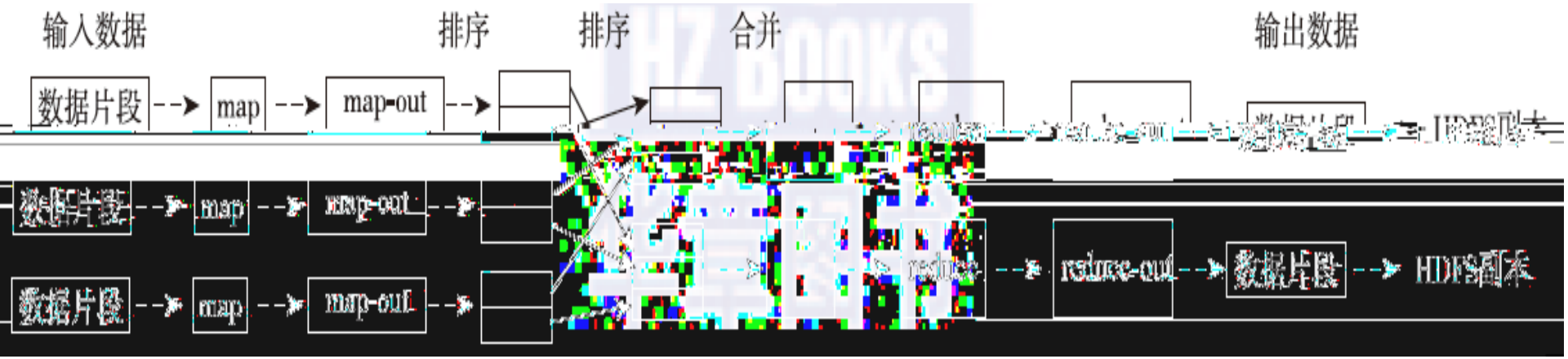
Hadoop与分布式开发

Hadoop上的并行应用程序开发是基于MapReduce 编程框架的。MapReduce 编程模型的原理是：利用一个输入的key/value 对集合来产生一个输出的key/value 对集合。MapReduce库的用户用两个函数来表达这个计算：Map 和Reduce。

用户自定义的map函数接收一个输入的key/value 对，然后产生一个中间key/value 对的集合。MapReduce 把所有具有相同key 值的value 集合在一起，然后传递给reduce 函数。用户自定义的reduce 函数接收key 和相关的value 集合。reduce 函数合并这些value 值，形成一个较小的value 集合。一般来说，每次reduce 函数调用只产生0 或1 个输出的value值。通常我们通过一个迭代器把中间的value 值提供给reduce 函数，这样就可以处理无法全部放入内存中的大量的value 值集合了。

Hadoop与分布式开发

下图是MapReduce 的数据流图，这个过程简而言之就是将大数据集分解为成百上千个小数据集，每个（或若干个）数据集分别由集群中的一个节点（一般就是一台普通的计算机）进行处理并生成中间结果，然后这些中间结果又由大量的节点合并，形成最终结果。图8-4也指出了MapReduce 框架下并行程序中的三个主要函数：map、reduce、main。在这个结构中，需要用户完成的工作仅仅是根据任务编写map和reduce 两个函数。



Hadoop与分布式开发

MapReduce 计算模型非常适合在大量计算机组成的大规模集群上并行运行。每一个 map 任务和每一个reduce 任务均可以同时运行于一个单独的计算节点上，可想而知，其运算效率是很高的，那么这样的并行计算是如何做到的呢？

1. 数据分布存储
2. 分布式并行计算
3. 本地计算
4. 任务粒度
5. 数据分割（Partition）
6. 数据合并（Combine）
7. Reduce
8. 任务管道

Hadoop应用案例

随着企业的数据量的迅速增长，存储和处理大规模数据已成为企业的迫切需求。Hadoop作为开源的云计算平台，已引起了学术界和企业的普遍兴趣。

在学术方面，Hadoop 得到了各科研院所的广泛关注，多所著名大学加入到Hadoop 集群的研究中来。
在商业方面，Hadoop 技术已经在互联网领域得到了广泛的应用。互联网公司往往需要存储海量的数据并对其进行处理，而这正是Hadoop 的强项。如Facebook 使用Hadoop 存储内部的日志拷贝，以及数据挖掘和日志统计；Yahoo ！利用Hadoop 支持广告系统并处理网页搜索；Twitter 则使用Hadoop 存储微博数据、日志文件和其他中间数据等。在国内，Hadoop同样也得到了许多公司的青睐，如百度主要将Hadoop 应用于日志分析和网页数据库的数据挖掘；阿里巴巴则将Hadoop 用于商业数据的排序和搜索引擎的优化等。

Hadoop平台上的海量数据排序

Yahoo! 研究人员使用Hadoop 完成了Jim Gray 基准排序，此排序包含许多相关的基准，每个基准都有自己的规则。所有的排序基准都是通过测量不同记录的排序时间来制定的，每个记录为100 字节，其中前面的10 字节是键，剩余的部分是数值。MinuteSort 是比较在一分钟内所排序的数据量大小，GraySort 是比较在对大规模数据（至少100TB）进行排序时的排序速率（TBs/minute）。

Yahoo! 的研究人员使用Hadoop 排列1TB 数据用时62 秒，排列1PB 数据用时16.25 个小时，具体如下表所示，它获得了Daytona 类GraySort 和MinuteSort 级别的冠军。

数据大小（Bytes）	节点数	副本数	时间
500 000 000 000	1406	1	59秒
1 000 000 000 000	1460	1	62秒
100 000 000 000 000	3452	2	173分钟
1000 000 000 000 000	3658	2	957分钟

本章小结

本章首先介绍了Hadoop分布式计算平台：它是由Apache 软件基金会开发的一个开源分布式计算平台。以Hadoop分布式文件系统（HDFS）和MapReduce 为核心的Hadoop 为用户提供了系统底层细节透明的分布式基础架构。由于Hadoop 拥有可计量、成本低、高效、可信等突出特点，基于Hadoop 的应用已经遍地开花，尤其是在互联网领域。

本章接下来介绍了Hadoop 项目及其结构，现在Hadoop 已经发展成为一个包含多个子项目的集合，被用于分布式计算，虽然Hadoop 的核心是Hadoop分布式文件系统和MapReduce，但Hadoop下的Avro、Hive、HBase 等子项目提供了互补性服务或在核心层之上提供了更高层的服务；接下来简要介绍了以HDFS 和MapReduce为核心的Hadoop 体系结构。

最后，介绍了Hadoop的典型应用案例以及在Hadoop平台上的海量数据排序。

主讲教师和助教

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

助教：赖明星

单位：厦门大学计算机科学系数据库实验室2011级硕士研究生（导师：林子雨）
E-mail: mingxinglai@gmail.com
个人主页：http://mingxinglai.com

欢迎访问《大数据技术基础》2013班级网站：http://dblab.xmu.edu.cn/node/423
本讲义PPT存在配套教材《大数据技术基础》，请到上面网站下载。

Department of Computer Science, Xiamen University, Sep, 2013