大数据创新人才培养系列

Spark 编程基础

Scala 版

SPARK PROGRAMMING
(SCALA EDITION)

◎ 林子雨 赖永炫 陶继平 编著

名校名师打造大数据领域精品力作
深入浅出，有效降低 Spark 技术学习门槛
资源全面，构建全方位一站式在线服务体系

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

大数据创新人才培养系列

Spark 编程基础

Scala 版

SPARK PROGRAMMING
(SCALA EDITION)

◎ 林子雨 赖永炫 陶继平 编著

人民邮电出版社
北京

图书在版编目（CIP）数据

Spark编程基础 : Scala版 / 林子雨, 赖永炫, 陶继平编著. -- 北京 : 人民邮电出版社, 2018.8
（大数据创新人才培养系列）
ISBN 978-7-115-48816-9

Ⅰ. ①S… Ⅱ. ①林… ②赖… ③陶… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第154480号

内 容 提 要

本书以 Scala 作为开发 Spark 应用程序的编程语言，系统地介绍了 Spark 编程的基础知识。全书共 8 章，内容包括大数据技术概述、Scala 语言基础、Spark 的设计与运行原理、Spark 环境搭建和使用方法、RDD 编程、Spark SQL、Spark Streaming 和 Spark MLlib。

本书每章都安排了入门级的编程实践操作，以便使读者能更好地学习和更牢固地掌握 Spark 编程方法。本书配套官网免费提供了全套的在线教学资源，包括讲义 PPT、习题、源代码、软件、数据集、授课视频、上机实验指南等。

本书可以作为高等院校计算机、软件工程、数据科学与大数据技术等专业的进阶级大数据课程教材，用于指导 Spark 编程实践，也可供相关技术人员参考。

◆ 编　　著　林子雨　赖永炫　陶继平
责任编辑　邹文波
责任印制　沈　蓉　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路 11 号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京隆昌伟业印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：16　　2018 年 8 月第 1 版
字数：428 千字　　2018 年 8 月北京第 1 次印刷

定价：49.80 元

读者服务热线：(010)81055256　印装质量热线：(010)81055316
反盗版热线：(010)81055315

前言 FOREWORD

大数据时代的来临，给各行各业带来了深刻的变革。大数据像能源、原材料一样，已经成为提升国家和企业竞争力的关键要素，被称为“未来的新石油”。正如电力技术的应用引发了生产模式的变革一样，基于互联网技术而发展起来的大数据技术的应用，将会为人们的生产和生活带来颠覆性的影响。

目前，大数据技术正处于快速发展之中，不断有新的技术涌现，Hadoop 和 Spark 等技术成为其中的佼佼者。在 Spark 流行之前，Hadoop 俨然已成为大数据技术的事实标准，在企业中得到了广泛的应用，但其本身还存在诸多缺陷，最主要的是 MapReduce 计算模型延迟过高，无法胜任实时、快速计算的需求，因而只适用于离线批处理的应用场景。Spark 在设计上充分吸收借鉴了 MapReduce 的精髓并加以改进，同时，采用了先进的 DAG 执行引擎，以支持循环数据流与内存计算，因此，在性能上比 MapReduce 有了大幅度的提升，从而迅速获得了学术界和业界的广泛关注。作为大数据计算平台的后起之秀，Spark 在 2014 年打破了 Hadoop 保持的基准排序纪录，此后逐渐发展成为大数据领域最热门的大数据计算平台之一。

随着大数据在企业应用的不断深化，企业对大数据人才的需求日益增长。为了有效地满足不断增长的大数据人才需求，国内高校从 2016 年开始设立“数据科学与大数据技术专业”，着力培养数据科学与工程领域的复合型高技术人才。课程体系的建设和课程教材的创作，是高校大数据专业建设的核心环节。

厦门大学数据库实验室在大数据教学领域辛勤耕耘、开拓创新，成为国内高校大数据教学资源的有力贡献者。实验室在积极践行 O2O 大数据教学理念的同时，提出了“以平台化思维构建全国高校大数据课程公共服务体系”的全新服务理念，成为推进国内高校大数据教学不断向前发展的一支重要力量，在全国高校之中形成了广泛的影响。2015 年 7 月，实验室编写出版了国内高校第一本系统性介绍大数据知识的专业教材——《大数据技术原理与应用》，受到了广泛的好评，目前已经成为国内众多高校的入门级大数据课程的开课教材。同时，实验室建设了国内高校首个大数据课程公共服务平台（网址：http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/），为全国高校教师和学生提供大数据教学资源一站式“免费”在线服务，包括课程教材、讲义 PPT、课程习题、实验指南、学习指南、备课指南、授课视频和技术资料等，自 2013 年 5 月建设以来，定位明确，进展顺利，目前平台每年访问量超过 100 万次，成为全国高校大数据教学的知名品牌。

《大数据技术原理与应用》定位为入门级大数据教材，以“构建知识体系、阐明基本原理、开展初级实践、了解相关应用”为原则，旨在为读者搭建起通向大数据知识空间的桥梁和纽带，为读者在大数据领域深耕细作奠定基础、指明方向。高校在开设入门级课程以后，可以根据自己的实际情况，开设进阶级的大数据课程，继续深化对大数据技术的学习，而 Spark 是目前比较理想的大数据进阶课程学习内容。因此，厦门大学数据库实验室组织具有丰富经验的一线大数据教师精心编写了本教材。

为了确保教材质量，在出版纸质图书之前，实验室已经于 2016 年 10 月通过实验室官网免费共享了简化版的 Spark 在线教程和相关教学资源，同时，该在线教程也已经用于厦门大学计算机科学系研究生的大数据课程教学，并成为全国高校大数据课程教师培训交流班的授课内容。实验室根据读者对在线 Spark 教程的大量反馈意见以及在教学实践中发现的问题，对 Spark 在线教程进行了多次修正和完善，这些前期准备工作，都为纸质图书的编著出版打下了坚实的基础。

本书共 8 章，详细介绍了 Spark 的环境搭建和基础编程方法。第 1 章介绍大数据关键技术，帮助读者对大数据技术形成总体性认识以及了解 Spark 在其中所扮演的角色；第 2 章介绍 Scala 语言基础知识，为学习基于 Scala 语言的 Spark 编程奠定基础；第 3 章介绍 Spark 的设计与运行原理；第 4 章介绍 Spark 的环境搭建和使用方法，为开展 Spark 编程实践铺平道路；第 5 章介绍 RDD 编程，包括 RDD 的创建、操作 API、持久化、分区以及键值对 RDD 等，这章知识是开展 Spark 高级编程的基础；第 6 章介绍 Spark 中用于结构化数据处理的组件 Spark SQL，包括 DataFrame 数据模型、创建方法和常用操作等；第 7 章介绍 Spark Streaming，这是一种构建在 Spark 上的流计算框架，可以满足对流式数据进行实时计算的需求；第 8 章介绍 Spark 的机器学习库 MLlib，包括 MLlib 的基本原理、算法、模型选择和超参数调整方法等。

本书面向高校计算机、软件工程、数据科学与大数据技术等专业的学生，可以作为专业必修课或选修课教材。本书由林子雨、赖永炫和陶继平执笔，其中，林子雨负责全书规划、统稿、校对和在线资源创作，并撰写第 1、3、5、6、7 章的内容，赖永炫负责撰写第 8 章的内容，陶

继平负责撰写第 2、4 章的内容。在撰写过程中，厦门大学计算机科学系硕士研究生阮榕城、薛倩、魏亮、曾冠华、程璐、林哲等做了大量的辅助性工作，在此，向这些同学的辛勤工作表示衷心的感谢。同时，感谢夏小云老师在书稿校对过程中的辛勤付出。

本书配套的官方网站是 http://dblab.xmu.edu.cn/post/spark/，免费提供全部配套资源的在线浏览和下载，并接受错误反馈和发布勘误信息。同时，Spark 作为大数据进阶课程，在学习过程中会涉及大量相关的大数据基础知识以及各种大数据软件的安装和使用方法，因此，推荐读者访问厦门大学数据库实验室建设的国内高校首个大数据课程公共服务平台（http://dblab.xmu.edu.cn/post/bigdata-teaching-platform/），来获得必要的辅助学习内容。

本书在撰写过程中，参考了大量的网络资料和相关书籍，对 Spark 技术进行了系统梳理，有选择性地把一些重要知识纳入本书。由于笔者能力有限，本书难免存在不足之处，望广大读者不吝赐教。

林子雨

厦门大学计算机科学系数据库实验室

2018 年 1 月

目 录 CONTENTS

01 第1章 大数据技术概述

大数据时代的来临，给各行各业带来了深刻的变革。大数据像能源、原材料一样，已经成为提升国家和企业竞争力的关键要素，被称为“未来的新石油”。正如电力技术的应用引发了生产模式的变革一样，基于互联网技术而发展起来的大数据应用，将会对人们的生产和生活产生颠覆性的影响。

本章首先介绍大数据的概念与关键技术，然后重点介绍有代表性的大数据技术，包括 Hadoop、Spark、Flink、Beam 等，最后探讨本教程编程语言的选择，并给出与本教材配套的相关在线资源。

1.1 大数据的概念与关键技术

随着大数据时代的到来，“大数据”已经成为互联网信息技术行业的流行词汇。本节介绍大数据的概念与关键技术。

1.1.1 大数据的概念

关于“什么是大数据”这个问题，学术界和业界比较认可关于大数据的“4V”说法。大数据的4个“V”，或者说是大数据的4个特点，包含4个层面：数据量大（Volume）、数据类型繁多（Variety）、处理速度快（Velocity）和价值密度低（Value）。

（1）数据量大。根据著名咨询机构IDC（Internet Data Center）做出的估测，人类社会产生的数据一直都在以每年50%的速度增长，这被称为“大数据摩尔定律”。这意味着，人类在最近两年产生的数据量相当于之前产生的全部数据量之和。预计到2020年，全球将总共拥有35ZB的数据量，数据量将增长到2010年数据的近30倍。

（2）数据类型繁多。大数据的数据类型丰富，包括结构化数据和非结构化数据，其中，前者占10%左右，主要是指存储在关系数据库中的数据，后者占90%左右，种类繁多，主要包括邮件、音频、视频、微信、微博、位置信息、链接信息、手机呼叫信息、网络日志等。

（3）处理速度快。大数据时代的很多应用，都需要基于快速生成的数据给出实时分析结果，用于指导生产和生活实践，因此，数据处理和分析的速度通常要达到秒级响应，这一点和传统的数据挖掘技术有着本质的不同，后者通常不要求给出实时分析结果。

（4）价值密度低。大数据价值密度却远远低于传统关系数据库中已经有的那些数据，在大数据时代，很多有价值的信息都是分散在海量数据中的。

1.1.2 大数据关键技术

大数据的基本处理流程，主要包括数据采集、存储管理、处理分析、结果呈现等环节。因此，从数据分析全流程的角度来看，大数据技术主要包括数据采集与预处理、数据存储和管理、数据处理与分析、数据可视化、数据安全和隐私保护等几个层面的内容（具体如表1-1所示）。

表1-1　　大数据技术的不同层面及其功能

技术层面	功能
数据采集与预处理	利用ETL（Extraction Transformation Loading）工具将分布的、异构数据源中的数据，如关系数据、平面数据文件等，抽取到临时中间层后进行清洗、转换、集成，最后加载到数据仓库或数据集市中，成为联机分析处理、数据挖掘的基础；也可以利用日志采集工具（如Flume、Kafka等）把实时采集的数据作为流计算系统的输入，进行实时处理分析
数据存储和管理	利用分布式文件系统、数据仓库、关系数据库、NoSQL数据库、云数据库等，实现对结构化、半结构化和非结构化海量数据的存储和管理
数据处理与分析	利用分布式并行编程模型和计算框架，结合机器学习和数据挖掘算法，实现对海量数据的处理和分析
数据可视化	对分析结果进行可视化呈现，帮助人们更好地理解数据、分析数据
数据安全和隐私保护	在从大数据中挖掘潜在的巨大商业价值和学术价值的同时，构建隐私数据保护体系和数据安全体系，有效保护个人隐私和数据安全

此外，大数据技术及其代表性软件种类繁多，不同的技术都有其适用和不适用的场景。总体而言，不同的企业应用场景，都对应着不同的大数据计算模式，根据不同的大数据计算模式，可以选择相应的大数据计算产品，具体如表 1-2 所示。

表 1-2　　大数据计算模式及其代表产品

大数据计算模式	解决问题	代表产品
批处理计算	针对大规模数据的批量处理	MapReduce、Spark 等
流计算	针对流数据的实时计算	Storm、S4、Flume、Streams、Puma、DStream、Super Mario、银河流数据处理平台等
图计算	针对大规模图结构数据的处理	Pregel、GraphX、Giraph、PowerGraph、Hama、GoldenOrb 等
查询分析计算	大规模数据的存储管理和查询分析	Dremel、Hive、Cassandra、Impala 等

批处理计算主要解决针对大规模数据的批量处理，也是我们日常数据分析工作中非常常见的一类数据处理需求。比如，爬虫程序把大量网页抓取过来存储到数据库中以后，可以使用 MapReduce 对这些网页数据进行批量处理，生成索引，加快搜索引擎的查询速度。代表性的批处理框架包括 MapReduce、Spark 等。

流计算主要是实时处理来自不同数据源的、连续到达的流数据，经过实时分析处理，给出有价值的分析结果。比如，用户在访问淘宝网等电子商务网站时，用户在网页中的每次点击的相关信息（比如选取了什么商品）都会像水流一样实时传播到大数据分析平台，平台采用流计算技术对这些数据进行实时处理分析，构建用户“画像”，为其推荐可能感兴趣的其他相关商品。代表性的流计算框架包括 Twitter Storm、Yahoo! S4 等。Twitter Storm 是一个免费、开源的分布式实时计算系统，Storm 对于实时计算的意义类似于 Hadoop 对于批处理的意义，Storm 可以简单、高效、可靠地处理流数据，并支持多种编程语言。Storm 框架可以方便地与数据库系统进行整合，从而开发出强大的实时计算系统。Storm 可用于许多领域中，如实时分析、在线机器学习、持续计算、远程 RPC、数据提取加载转换等。由于 Storm 具有可扩展、高容错性、能可靠地处理消息等特点，目前已经被广泛应用于流计算应用中。

在大数据时代，许多大数据都是以大规模图或网络的形式呈现，如社交网络、传染病传播途径、交通事故对路网的影响等。此外，许多非图结构的大数据，也常常会被转换为图模型后再进行处理分析。图计算软件是专门针对图结构数据开发的，在处理大规模图结构数据时可以获得很好的性能。谷歌公司的 Pregel 是一种基于 BSP 模型实现的图计算框架。为了解决大型图的分布式计算问题，Pregel 搭建了一套可扩展的、有容错机制的平台，该平台提供了一套非常灵活的 API，可以描述各种各样的图计算。Pregel 作为分布式图计算的计算框架，主要用于图遍历、最短路径、PageRank 计算等。

查询分析计算也是一种在企业中常见的应用场景，主要是面向大规模数据的存储管理和查询分析，用户一般只需要输入查询语句（如 SQL），就可以快速得到相关的查询结果。典型的查询分析计算产品包括 Dremel、Hive、Cassandra、Impala 等。其中，Dremel 是一种可扩展的、交互式的实时查询系统，用于只读嵌套数据的分析。通过结合多级树状执行过程和列式数据结构，它能做到几秒内完成对万亿张表的聚合查询。系统可以扩展到成千上万的 CPU 上，满足谷歌上万用户操作 PB 级的数据，并且可以在 2～3 秒完成 PB 级别数据的查询。Hive 是一个构建于 Hadoop 顶层的数据仓库工具，允许用户输入 SQL 语句进行查询。Hive 在某种程度上可以看作是用户编程接口，其本身并不存

储和处理数据，而是依赖 HDFS 来存储数据，依赖 MapReduce 来处理数据。Hive 作为现有比较流行的数据仓库分析工具之一，得到了广泛的应用，但是由于 Hive 采用 MapReduce 来完成批量数据处理，因此，实时性不好，查询延迟较高。Impala 作为新一代开源大数据分析引擎，支持实时计算，它提供了与 Hive 类似的功能，通过 SQL 语句能查询存储在 Hadoop 的 HDFS 和 HBase 上的 PB 级别海量数据，并在性能上比 Hive 高出 3～30 倍。

1.2 代表性大数据技术

大数据技术的发展步伐很快，不断有新的技术涌现，这里着重介绍几种目前市场上具有代表性的一些大数据技术，包括 Hadoop、Spark、Flink、Beam 等。

1.2.1 Hadoop

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计算架构。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（Hadoop Distributed File System，HDFS）和 MapReduce。借助于 Hadoop，程序员可以轻松地编写分布式并行程序，将其运行在廉价的计算机集群上，完成海量数据的存储与计算。经过多年的发展，Hadoop 生态系统不断完善和成熟，目前已经包含多个子项目（见图 1-1）。除了核心的 HDFS 和 MapReduce 以外，Hadoop 生态系统还包括 YARN、Zookeeper、HBase、Hive、Pig、Mahout、Sqoop、Flume、Ambari 等功能组件。

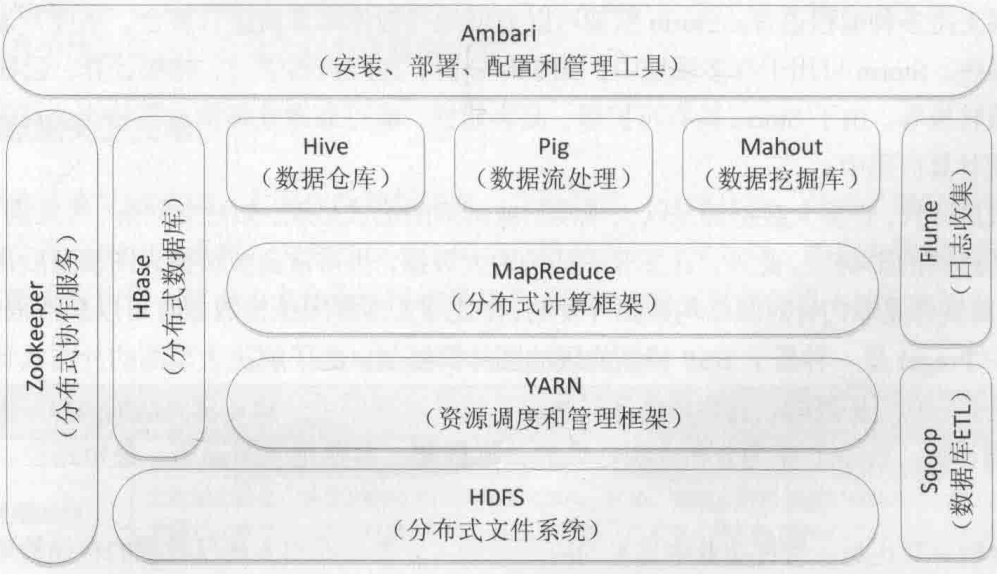

图 1-1 Hadoop 生态系统

这里简要介绍一下这些组件的功能，要了解 Hadoop 的更多细节内容，可以访问本教材官网，学习《大数据技术原理与应用》在线视频的内容。

1. HDFS

Hadoop 分布式文件系统 HDFS 是针对谷歌分布式文件系统（Google File System，GFS）的开源

实现，它是 Hadoop 两大核心组成部分之一，提供了在廉价服务器集群中进行大规模分布式文件存储的能力。HDFS 具有很好的容错能力，并且兼容廉价的硬件设备，因此，可以以较低的成本利用现有机器实现大流量和大数据量的读写。

HDFS 采用了主从（Master/Slave）结构模型，一个 HDFS 集群包括一个名称节点和若干个数据节点（见图 1-2）。名称节点作为中心服务器，负责管理文件系统的命名空间及客户端对文件的访问。集群中的数据节点一般是一个节点运行一个数据节点进程，负责处理文件系统客户端的读/写请求，在名称节点的统一调度下进行数据块的创建、删除和复制等操作。

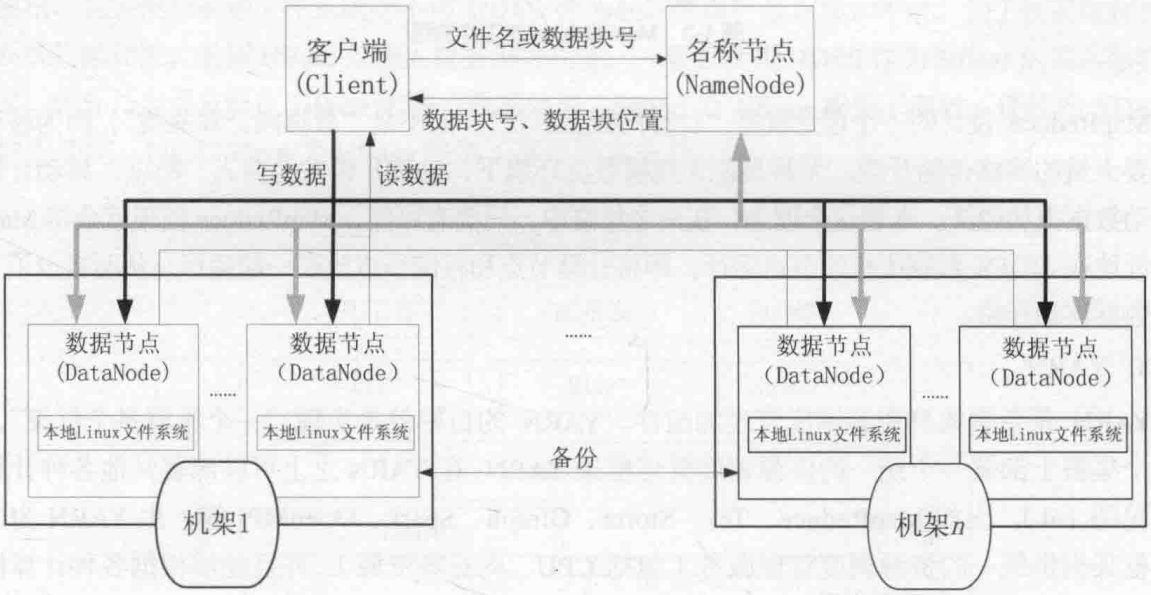

图 1-2 HDFS 的体系结构

用户在使用 HDFS 时，仍然可以像在普通文件系统中那样，使用文件名去存储和访问文件。实际上，在系统内部，一个文件会被切分成若干个数据块，这些数据块被分布存储到若干个数据节点上。当客户端需要访问一个文件时，首先把文件名发送给名称节点，名称节点根据文件名找到对应的数据块（一个文件可能包括多个数据块），再根据每个数据块信息找到实际存储各个数据块的数据节点的位置，并把数据节点位置发送给客户端，最后，客户端直接访问这些数据节点获取数据。在整个访问过程中，名称节点并不参与数据的传输。这种设计方式，使得一个文件的数据能够在不同的数据节点上实现并发访问，大大提高了数据的访问速度。

2. MapReduce

MapReduce 是一种分布式并行编程模型，用于大规模数据集（大于 1TB）的并行运算，它将复杂的、运行于大规模集群上的并行计算过程高度抽象到两个函数：Map 和 Reduce。MapReduce 极大方便了分布式编程工作，编程人员在不会分布式并行编程的情况下，也可以很容易将自己的程序运行在分布式系统上，完成海量数据集的计算。

在 MapReduce 中（见图 1-3），一个存储在分布式文件系统中的大规模数据集，会被切分成许多独立的小数据块，这些小数据块可以被多个 Map 任务并行处理。MapReduce 框架会为每个 Map 任务输入一个数据子集，Map 任务生成的结果会继续作为 Reduce 任务的输入，最终由 Reduce 任务输出最后结果，并写入分布式文件系统。

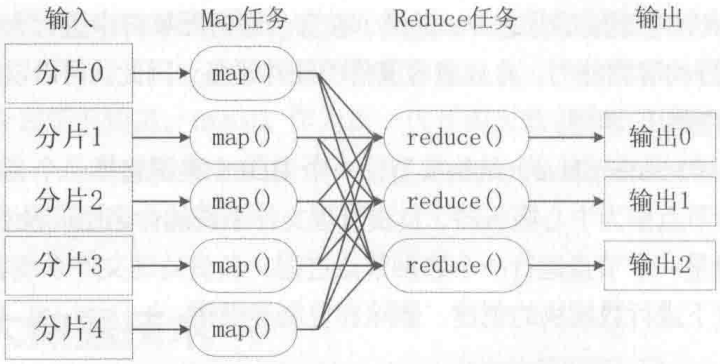

图 1-3　MapReduce 的工作流程

MapReduce 设计的一个理念就是“计算向数据靠拢”，而不是“数据向计算靠拢”，因为移动数据需要大量的网络传输开销，尤其是在大规模数据环境下，这种开销尤为惊人，所以，移动计算要比移动数据更加经济。本着这个理念，在一个集群中，只要有可能，MapReduce 框架就会将 Map 程序就近地在 HDFS 数据所在的节点运行，即将计算节点和存储节点放在一起运行，从而减少了节点间的数据移动开销。

3. YARN

YARN 是负责集群资源调度管理的组件。YARN 的目标就是实现“一个集群多个框架”，即在一个集群上部署一个统一的资源调度管理框架 YARN，在 YARN 之上可以部署其他各种计算框架（见图 1-4），比如 MapReduce、Tez、Storm、Giraph、Spark、OpenMPI 等，由 YARN 为这些计算框架提供统一的资源调度管理服务（包括 CPU、内存等资源），并且能够根据各种计算框架的负载需求，调整各自占用的资源，实现集群资源共享和资源弹性收缩。通过这种方式，可以实现一个集群上的不同应用负载混搭，有效提高了集群的利用率，同时，不同计算框架可以共享底层存储，在一个集群上集成多个数据集，使用多个计算框架来访问这些数据集，从而避免了数据集跨集群移动，最后，这种部署方式也大大降低了企业运维成本。目前，可以运行在 YARN 之上的计算框架包括离线批处理框架 MapReduce、内存计算框架 Spark、流计算框架 Storm 和 DAG 计算框架 Tez 等。和 YARN 一样提供类似功能的其他资源管理调度框架还包括 Mesos、Torca、Corona、Borg 等。

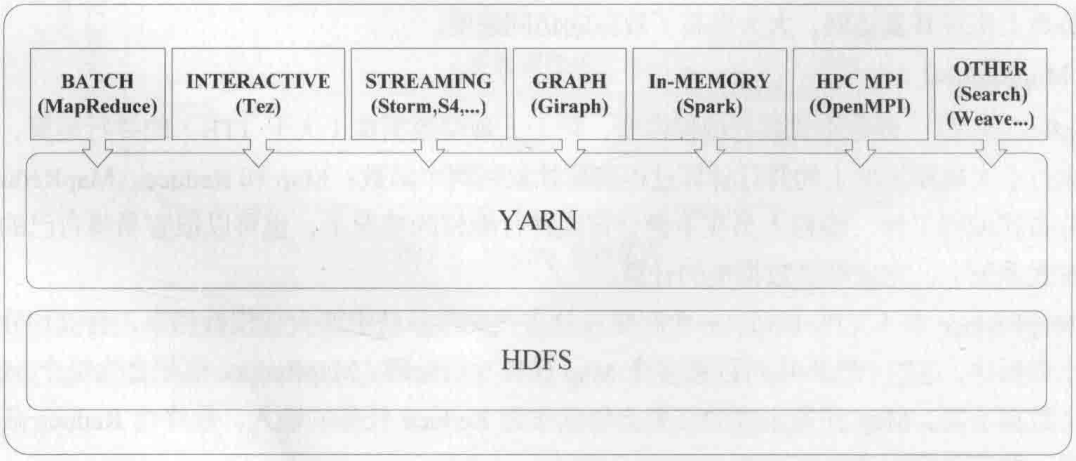

图 1-4　在 YARN 上部署各种计算框架

4. HBase

HBase 是针对谷歌 BigTable 的开源实现，是一个高可靠、高性能、面向列、可伸缩的分布式数据库，主要用来存储非结构化和半结构化的松散数据。HBase 可以支持超大规模数据存储，它可以通过水平扩展的方式，利用廉价计算机集群处理由超过 10 亿行元素和数百万列元素组成的数据表。

图 1-5 描述了 Hadoop 生态系统中 HBase 与其他部分的关系。HBase 利用 MapReduce 来处理 HBase 中的海量数据，实现高性能计算；利用 Zookeeper 作为协同服务，实现稳定服务和失败恢复；使用 HDFS 作为高可靠的底层存储，利用廉价集群提供海量数据存储能力，当然，HBase 也可以在单机模式下使用，直接使用本地文件系统而不用 HDFS 作为底层数据存储方式，不过，为了提高数据可靠性和系统的健壮性，发挥 HBase 处理大量数据等功能，一般都使用 HDFS 作为 HBase 的底层数据存储方式。此外，为了方便在 HBase 上进行数据处理，Sqoop 为 HBase 提供了高效、便捷的 RDBMS 数据导入功能，Pig 和 Hive 为 HBase 提供了高层语言支持。

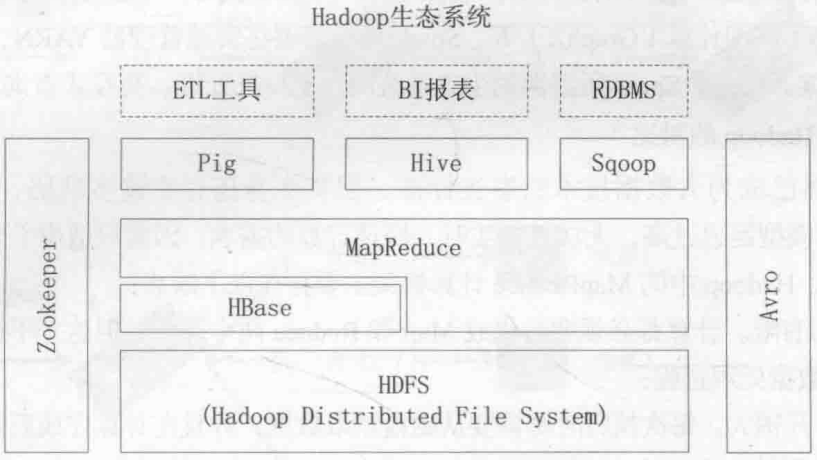

图 1-5　Hadoop 生态系统中 HBase 与其他部分的关系

5. Hive

Hive 是一个基于 Hadoop 的数据仓库工具，可以用于对存储在 Hadoop 文件中的数据集进行数据整理、特殊查询和分析处理。Hive 的学习门槛比较低，因为它提供了类似于关系数据库 SQL 语言的查询语言——HiveQL，可以通过 HiveQL 语句快速实现简单的 MapReduce 统计，Hive 自身可以自动将 HiveQL 语句快速转换成 MapReduce 任务进行运行，而不必开发专门的 MapReduce 应用程序，因而十分适合数据仓库的统计分析。

6. Flume

Flume 是 Cloudera 公司开发的一个高可用的、高可靠的、分布式的海量日志采集、聚合和传输系统。Flume 支持在日志系统中定制各类数据发送方，用于收集数据；同时，Flume 提供对数据进行简单处理，并写到各种数据接收方的能力。

7. Sqoop

Sqoop 是 SQL-to-Hadoop 的缩写，主要用来在 Hadoop 和关系数据库之间交换数据，可以改进数据的互操作性。通过 Sqoop，可以方便地将数据从 MySQL、Oracle、PostgreSQL 等关系数据库中导入 Hadoop（比如导入到 HDFS、HBase 或 Hive 中），或者将数据从 Hadoop 导出到关系数据库，使得