大数据人才培养规划教材

以解决实际问题为**学习目标**
以实战案例贯穿为**学习手段**

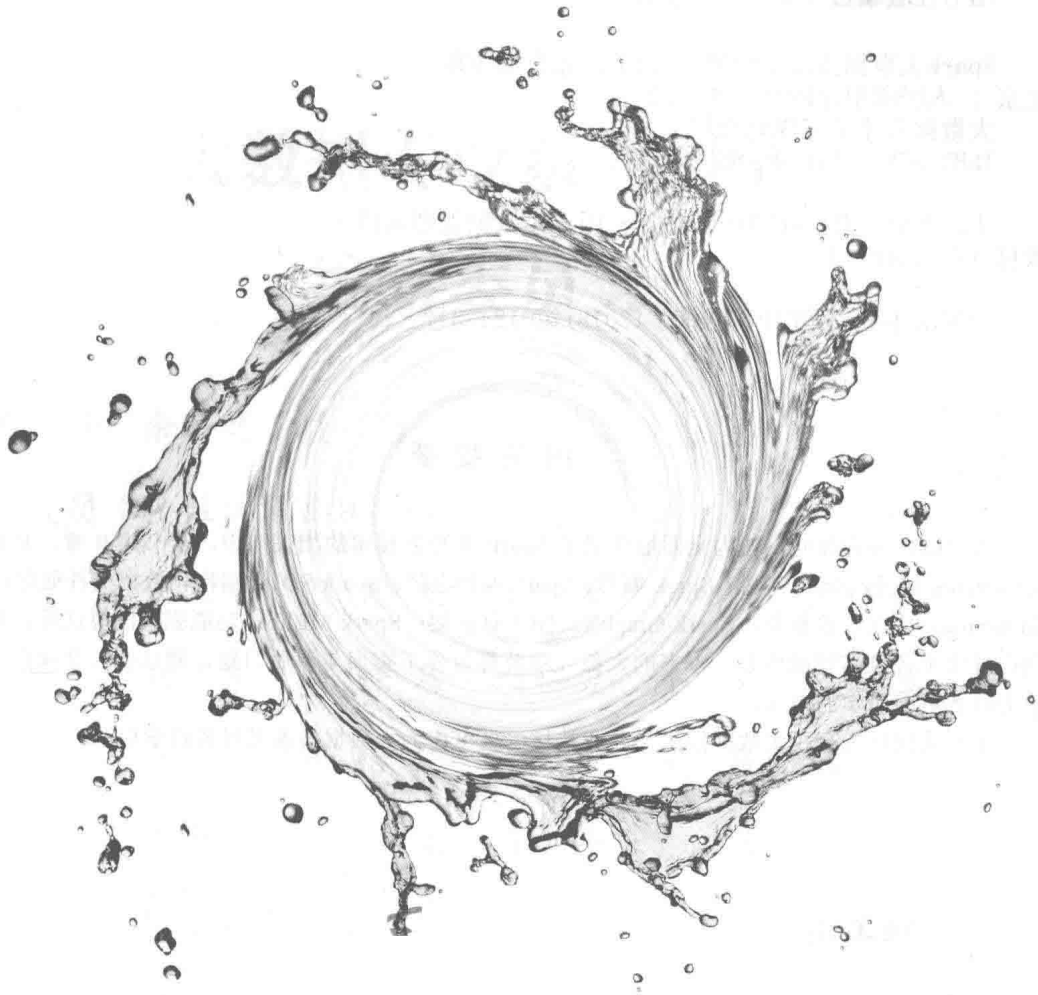
Spark
大数据技术与应用

Spark Big Data Technology and Application

肖芳 张良均 ◉ 主编
汪作文 胡大威 樊哲 ◉ 副主编

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

大数据人才培养规划教材



Spark 大数据技术与应用

Spark Big Application

肖芳 张良均 ◉ 主编
汪作文 胡大威 樊哲 ◉ 副主编

人 民 邮 电 出 版 社
北 京

图书在版编目（CIP）数据

Spark大数据技术与应用 / 肖芳，张良均主编. —
北京 : 人民邮电出版社，2018.2
大数据人才培养规划教材
ISBN 978-7-115-46488-0

Ⅰ. ①S… Ⅱ. ①肖… ②张… Ⅲ. ①数据处理软件－
教材 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2018)第012433号

内 容 提 要

本书以任务为导向，较为全面地介绍了 Spark 大数据技术的相关知识。全书共 9 章，具体内容包括 Spark 概述；Scala 基础；Spark 编程；Spark 编程进阶；Spark SQL：结构化数据文件处理；Spark Streaming：实时计算框架；Spark GraphX：图计算框架；Spark MLlib：功能强大的算法库；项目案例：餐饮平台菜品智能推荐。本书的大部分章节都包含了实训与课后习题，通过练习和操作实践，帮助读者巩固所学的内容。

本书可以作为高校大数据技术类专业教材，也可作为大数据技术爱好者自学用书。

◆ 主　　编　肖　芳　张良均
副 主 编　汪作文　胡大威　樊　哲
责任编辑　左仲海
责任印制　马振武

◆ 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
中国铁道出版社印刷厂印刷

◆ 开本：787×1092　1/16
印张：17.75　　2018 年 2 月第 1 版
字数：406 千字　　2018 年 2 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
广告经营许可证：京东工商广登字 20170147 号

大数据专业系列图书
编写委员会

序 PREFACE

随着大数据时代的到来，移动互联网络和智能手机迅速普及，多种形态的移动互联应用蓬勃发展，电子商务、云计算、互联网金融、物联网等不断渗透并重塑传统产业，大数据当之无愧地成为了新的产业革命核心。

未来 5～10 年，我国大数据产业将会是一个飞速发展时期，社会对大数据相关专业人才有着巨大的需求。目前，国内各大高校都在争相设立或准备设立大数据相关专业，以适应地方产业发展对战略性新兴产业的人才需求。

人才培养离不开教材，大数据专业是 2016 年才获批的新专业，目前还没有成套的系列教材，已有教材也存在企业案例缺失等急需解决的问题。由广州泰迪智能科技有限公司和人民邮电出版社策划，校企联合编写的这套图书，尤如大旱中的甘露，可以有效解决高校大数据相关专业教材紧缺的困境。

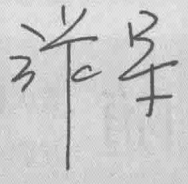
实践教学是在一定的理论指导下，通过引导学习者的实践活动，从而传承实践知识、形成技能、发展实践能力、提高综合素质的教学活动。目前，高校教学体系的设置有诸多限制因素，过多地偏向理论教学，课程设置与企业实际应用切合度不高，学生无法把理论转化为实践应用技能。课程内容设置方面看似繁多又各自为“政”，课程冗余、缺漏，体系不健全。本套图书的第一大特点就是注重学生的实践能力培养，根据高校实践教学中的痛点，首次提出“鱼骨教学法”的概念。以企业真实需求为导向，学生学习技能紧紧围绕企业实际应用需求，将学生需掌握的理论知识，通过企业案例的形式进行衔接，达到知行合一、以用促学的目的。

大数据专业应该以大数据技术应用为核心，紧紧围绕大数据应用闭环的流程进行教学，才能够使学生从宏观上理解大数据技术在行业中的具体应用场景及应用方法。高校现有的大数据课程集中在如何进行数据处理、建模分析、调整参数，使得模型的结果更加准确。但是，完整的大数据应用却是一个容易被忽视的部分。本套图书的第二大特点就是围绕大数据应用的整个流程，从数据采集、数据迁移、数据存储、数据

分析与挖掘，最终到数据可视化，覆盖完整的大数据应用流程，涵盖企业大数据应用中的各个环节，符合企业大数据应用真实场景。

我很高兴看到这套书的出版，也希望这套书能给更多的高校师生带来教学上的便利，帮助读者尽快掌握本领，成为有用之才！

教育部长江学者特聘教授
国家杰出青年基金获得者
IEEE Fellow
华南理工大学计算机与工程学院院长



2017 年 12 月

前 言 FOREWORD

为什么要学习 Spark

随着大数据时代的到来，各行各业的工作者都迫切需要更好更快的数据计算与分析工具。2009 年，Spark 应运而生，在很短的时间里就崭露头角，受到了业界的广泛肯定与欢迎，如今已是 Apache 软件基金会下的顶级开源项目之一。相较于曾经引爆大数据产业革命的 Hadoop MapReduce 框架，Spark 带来的改进更加令人欢欣鼓舞。首先，基于内存计算的 Spark 速度更快，减少了迭代计算时的 IO 开销，而且支持交互性使用。其次，Spark 丰富的 API 提供了更强大的易用性，它支持使用 Scala、Java、Python 与 R 语言进行编程，有助于开发者轻松构建并行的应用程序。而且，Spark 支持多种运行模式，既可以运行于独立的集群模式或 Hadoop 集群模式中，也可以运行于 Amazon EC2 等云环境中，并且可以访问 HDFS、HBase、Hive、MySQL 等多种数据源。最后，Spark 是一个通用的引擎，支持各种各样的运算模式，除了传统批处理应用外，还包括了 SQL 查询、流式计算、机器学习、图计算等应用，因此它能够更加灵活地满足不同场景下的应用需求，将多个组件无缝整合在同一个应用中。因为以上的这些优点，使得 Spark 成为学习大数据技术的一个绝佳的起点。

如何带领读者从零基础开始学习 Spark 大数据技术，并能够理论结合实践，运用 Spark 相关技术知识解决一些实际的业务需求，这正是本书要解决的问题。

本书特色

本书是定位于 Spark 大数据技术从入门到应用的简明系统教程，主要包括 Spark 基本原理与架构、集群安装配置、Scala 与 Spark 编程、Spark 代表组件、完整项目案例等精选内容。涉及的知识点简要精到，实践操作性强，使用本书能有效指导读者对 Spark 大数据技术的学习理解及开发应用。

本书采用了以任务为导向的教学模式，按照解决实际任务的工作流程路线，逐步展开学习相关的理论知识点，推导生成可行的解决方案，最后落实在任务实现环节。全书大部分章节紧扣任务需求展开，不堆积知识点，着重于解决思路的启发与解决方案的实施。通过从任务需求到实现这一完整工作流程的体验，有助于读者对 Spark 大数据技术真正的理解与掌握。

本书适用对象

- 开设有大数据相关课程的高校教师和学生

目前国内不少高校将大数据技术引入教学中，在计算机、数学、自动化、电子信息、金融等专业开设了与大数据技术相关的课程，但目前这一课程教学用的相关教材没有统一，或者使用的教材不利于课堂教学。本书提供了 Spark 相关技术的介绍、原理、实践、企业应用等，能有效指导高校教师和学生对 Spark 相关技术原理、技术实践有一定认识，为以后工作打下良好基础。

- 大数据开发技术人员

书中针对 Spark 大数据开发技术，由浅入深地做了系统性的介绍，并且针对每一模块有对应的动手实践，对于初级开发人员有较强的指导作用。

- 关注大数据技术的各行业技术人员

本书不仅对 Spark 大数据相关技术进行理论性的介绍及讲解，还包括了多个行业实践任务与大数据相关技术相结合的案例分析。各行业技术人员可以通过学习书中案例的解决思路与实现方法，尝试以新技术解决本行业中的相关问题。

代码下载及问题反馈

为方便读者实践与练习，书中提供全部实例的数据文件及源代码，读者可登录人民邮电出版社教育社区（www.ryjiaoyu.com）或“泰迪杯”全国数据挖掘挑战赛网站（www.tipdm.org/tj/1305.jhtml）下载。为方便广大教师授课，本书也提供了教学课件 PPT。如有需要可通过泰迪大数据挖掘微信公众号（TipDataMining）或者热线电话（40068-40020）进行在线咨询获取。

由于编者水平有限，编写时间仓促，书中难免出现一些疏漏和不足。如果您有更多的宝贵意见，欢迎发送邮件至邮箱 13560356095@qq.com，期待得到您真挚的反馈。同时，本书的内容更新将及时在“泰迪杯”全国数据挖掘挑战赛网站上发布，读者可以登录网站或关注泰迪大数据挖掘微信公众号（TipDataMining）查阅相关信息。

编　者

2017 年 10 月

目录 CONTENTS

第 1 章 Spark 概述 …… 1

任务 1.1 认识 Spark …… 1

1.1.1 Spark 的发展 …… 1

1.1.2 Spark 的特点 …… 2

1.1.3 Spark 生态圈 …… 4

1.1.4 Spark 的应用场景 …… 5

任务 1.2 搭建 Spark 环境 …… 5

1.2.1 搭建单机版环境 …… 6

1.2.2 搭建单机伪分布式环境 …… 6

1.2.3 搭建完全分布式环境 …… 7

任务 1.3 了解 Spark 运行架构与原理 …… 10

1.3.1 Spark 集群架构 …… 11

1.3.2 Spark 作业运行流程 …… 11

1.3.3 Spark 核心数据集 RDD …… 15

1.3.4 Spark 核心原理 …… 17

小结 …… 19

第 2 章 Scala 基础 …… 20

任务 2.1 Scala 的简介与安装 …… 21

2.1.1 Scala 简介 …… 21

2.1.2 Scala 特性 …… 21

2.1.3 Scala 的环境设置及安装 …… 21

2.1.4 运行 Scala …… 23

任务 2.2 定义函数识别号码类型 …… 24

2.2.1 数据类型 …… 24

2.2.2 常量和变量 …… 25

2.2.3 表达式 …… 26

2.2.4 数组 …… 27

2.2.5 函数 …… 29

2.2.6 任务实现 …… 31

任务 2.3 统计广州号码段数量 …… 32

2.3.1 if 判断 …… 32

2.3.2 循环 …… 33

2.3.3 任务实现 …… 34

任务 2.4 根据归属地对手机号码段分组 …… 34

2.4.1 List …… 35

2.4.2 Set …… 36

2.4.3 Map …… 37

2.4.4 元组 …… 38

2.4.5 函数组合器 …… 38

2.4.6 任务实现 …… 39

任务 2.5 编写手机号码归属地信息查询程序 …… 40

2.5.1 Scala 类 …… 40

2.5.2 Scala object …… 42

2.5.3 Scala 模式匹配 …… 42

2.5.4 Scala 读取文件 …… 44

2.5.5 任务实现 …… 44

小结 …… 45

实训 …… 45

实训 1 编写函数过滤文本中的回文单词 …… 45

实训 2 使用 Scala 编程输出杨辉三角 …… 46

课后习题 …… 46

第 3 章 Spark 编程 …… 48

任务 3.1 以学生成绩数据创建 RDD … 49

3.1.1 从内存中已有数据创建 RDD …… 50

3.1.2 从外部存储创建 RDD …… 51

3.1.3 任务实现 …… 52

任务 3.2 查询学生成绩表中的前 5 名 …… 52
3.2.1 使用 map 转换数据 …… 52
3.2.2 使用 sortBy()排序 …… 53
3.2.3 使用 collect()查询 …… 53
3.2.4 使用 flatMap 转换数据 …… 54
3.2.5 使用 take()方式查询某几个值 …… 54
3.2.6 任务实现 …… 55
任务 3.3 输出单科成绩为 100 分的学生 ID …… 55
3.3.1 使用 union()合并多个 RDD …… 56
3.3.2 使用 filter()进行过滤 …… 56
3.3.3 使用 distinct()进行去重 …… 56
3.3.4 简单的集合操作 …… 57
3.3.5 任务实现 …… 58
任务 3.4 输出每位学生所有科目的总成绩 …… 58
3.4.1 键值对 RDD 简介 …… 59
3.4.2 创建键值对 RDD …… 59
3.4.3 转换操作 keys 与 values …… 59
3.4.4 转换操作 reduceByKey() …… 60
3.4.5 转换操作 groupByKey() …… 60
3.4.6 任务实现 …… 60
任务 3.5 输出每位学生的平均成绩 …… 61
3.5.1 使用 join()连接两个 RDD …… 61
3.5.2 使用 zip 组合两个 RDD …… 63
3.5.3 使用 combineByKey 合并相同键的值 …… 63
3.5.4 使用 lookup 查找指定键的值 …… 64
3.5.5 任务实现 …… 64
任务 3.6 将汇总后的学生成绩存储为文本文件 …… 65
3.6.1 JSON 文件的读取与存储 …… 65
3.6.2 CSV 文件的读取与存储 …… 67
3.6.3 SquenceFile 的读取与存储 …… 69
3.6.4 文本文件的读取与存储 …… 70
3.6.5 任务实现 …… 71
小结 …… 72
实训 …… 72
实训 1 统计文本中性别为“男”的用户数 …… 73
实训 2 单词计数 …… 73
课后习题 …… 74
第 4 章 Spark 编程进阶 …… 76
任务 4.1 搭建开发环境 …… 77
4.1.1 下载与安装 IntelliJ IDEA …… 77
4.1.2 Scala 插件安装与使用 …… 79
4.1.3 配置 Spark 运行环境 …… 84
4.1.4 运行 Spark 程序 …… 85
任务 4.2 使用移动平均预测股票涨跌 …… 92
4.2.1 持久化（缓存） …… 93
4.2.2 数据分区 …… 94
4.2.3 计算价格波动幅度 …… 98
4.2.4 任务实现 …… 100
小结 …… 103
实训 …… 103
实训 竞赛网站访问日志分析 …… 104
课后习题 …… 104
第 5 章 Spark SQL：结构化数据文件处理 …… 107
任务 5.1 认识 Spark SQL …… 108
5.1.1 Spark SQL 简介 …… 108
5.1.2 Spark SQL CLI 配置 …… 109
5.1.3 Spark SQL 与 Shell 交互 …… 110
任务 5.2 掌握 DataFrame 基础操作 …… 111
5.2.1 创建 DataFrame 对象 …… 111
5.2.2 DataFrame 查看数据 …… 114
5.2.3 DataFrame 查询操作 …… 117
5.2.4 DataFrame 输出操作 …… 123
任务 5.3 探索分析法律服务网站数据 …… 125
5.3.1 获取数据 …… 125
5.3.2 网页类型分析 …… 126
5.3.3 点击次数分析 …… 131
5.3.4 网页排名分析 …… 133

小结 …… 135
实训 …… 135
实训 1 统计分析航空公司客户数据的空值以及异常值 …… 135
实训 2 统计分析某公司每年的产品销售量及销售额 …… 137
课后习题 …… 139
第 6 章 Spark Streaming：实时计算框架 …… 141
任务 6.1 初探 Spark Streaming …… 142
6.1.1 Spark Streaming 概述 …… 142
6.1.2 Spark Streaming 运行原理 …… 142
6.1.3 初步使用 Spark Streaming …… 143
任务 6.2 掌握 DStream 编程模型 …… 145
6.2.1 DStream 简介 …… 146
6.2.2 DStream 转换操作 …… 146
6.2.3 DStream 窗口操作 …… 148
6.2.4 DStream 输出操作 …… 151
任务 6.3 Spark Streaming 实时更新热门博文 …… 155
6.3.1 Spark Streaming 输入数据源 …… 155
6.3.2 Spark Streaming 计算网页热度 …… 158
6.3.3 网页热度输出 …… 158
6.3.4 任务实现 …… 159
小结 …… 161
实训 …… 161
实训 1 过滤打印包含单词 error 的记录 …… 162
实训 2 实时过滤歌曲播放次数超过 100 次的记录并存储在 HDFS 上 …… 162
课后习题 …… 162
第 7 章 Spark GraphX：图计算框架 …… 165
任务 7.1 认识 Spark GraphX …… 166
7.1.1 图的基本概念 …… 166
7.1.2 图计算的应用 …… 167
7.1.3 GraphX 的基础概念 …… 168
7.1.4 GraphX 的发展 …… 168
任务 7.2 了解 GraphX 常用 API …… 169
7.2.1 图的创建与存储 …… 169
7.2.2 数据查询与数据转换 …… 174
7.2.3 结构转换与关联聚合 …… 180
任务 7.3 构建信任网络并找出目标用户 …… 187
7.3.1 构建网站信任网络 …… 188
7.3.2 找出需要支付稿酬的用户 …… 188
7.3.3 找出进入热门榜的用户 …… 189
小结 …… 191
实训 …… 191
实训 1 使用 PageRank 算法完成网页排名 …… 191
实训 2 利用二度关系完成商品推荐 …… 192
课后习题 …… 194
第 8 章 Spark MLlib：功能强大的算法库 …… 196
任务 8.1 了解 MLlib 算法库 …… 197
8.1.1 机器学习简介 …… 197
8.1.2 MLlib 介绍 …… 198
任务 8.2 以 Logistic 回归实现用户分类 …… 212
8.2.1 分析思路 …… 212
8.2.2 数据处理 …… 213
8.2.3 MLlib 实现 Logistic 回归 …… 215
8.2.4 任务实现 …… 217
小结 …… 221
实训 …… 221
实训 1 通过 KMeans 定位商圈 …… 221
实训 2 朴素贝叶斯进行文本分类 …… 222
课后习题 …… 223
第 9 章 项目案例：餐饮平台菜品智能推荐 …… 226
任务 9.1 推荐方案设计 …… 227
9.1.1 用户数据分析 …… 227
9.1.2 常用推荐算法 …… 229

9.1.3 推荐流程设计 …… 231
任务 9.2 数据预处理 …… 232
9.2.1 原始数据探索分析 …… 233
9.2.2 异常数据处理 …… 237
9.2.3 数据变换处理 …… 237
9.2.4 数据集分割 …… 239
任务 9.3 建立推荐模型 …… 240
9.3.1 以基于用户的协同过滤算法建模 …… 240
9.3.2 以基于物品的协同过滤算法建模 …… 243
9.3.3 以基于 Spark ALS 的协同过滤算法建立模型 …… 246
9.3.4 推荐模型的评测 …… 251
任务 9.4 使用模型进行菜品推荐 …… 262
9.4.1 对某用户推荐 10 道新菜品 …… 262
9.4.2 对所有用户进行新菜品推荐 …… 267
小结 …… 272

第 1 章 Spark 概述

学习目标

（1）了解 Spark 的发展历史及特点。
（2）学会搭建 Spark 环境。
（3）了解 Spark 的运行架构与原理。

任务背景

大数据技术蓬勃发展，基于开源技术的 Hadoop 在行业中应用广泛。但是 Hadoop 本身还存在诸多缺陷，最主要的缺陷是其 MapReduce 计算模型延迟过高，无法胜任实时、快速计算的需求。Spark 的诞生弥补了 MapReduce 的缺陷。Spark 继承了 MapReduce 分布式计算的优点并改进了 MapReduce 明显的缺陷。Spark 拥有 Hadoop MapReduce 所具有的优点，但不同于MapReduce，Spark 的中间输出结果可以保存在内存中，从而大大减少了读写 HDFS 的次数，因此 Spark 能更好地适用数据挖掘与机器学习中需要迭代的算法。

任务 1.1 认识 Spark

任务描述

学习 Spark，首先应该了解 Spark 的发展及特点，本节的任务是带领读者走进 Spark，了解 Spark 的发展历史，理解 Spark 的特点，同时认识 Spark 的生态圈并了解 Spark 的应用场景。

1.1.1 Spark 的发展

图 1-1 展示了 Spark 的发展历史，对于一个具有相当技术门槛与复杂度的平台，Spark 从诞生到正式版本的成熟，经历的时间如此之短，让人感到惊诧。目前，Spark 已经成为 Apache 软件基金会旗下的顶级开源项目。下面是 Spark 的发展历程简述。

2009 年，Spark 诞生于伯克利大学 AMPLab，最初属于伯克利大学的研究性项目，实验室的研究人员之前基于 Hadoop MapReduce 工作，他们发现 MapReduce 对于迭代和交互式计算任务效率不高，因此他们研究的 Spark 主要为交互式查询和迭代算法设计，支持内存存储和高效的容错恢复。

2010 年，Spark 正式开源。

2013 年 6 月，Spark 成为 Apache 基金会的孵化器项目。

2014 年 2 月，仅仅经历 8 个月的时间，Spark 就成为 Apache 基金会的顶级项目。同时，大数据公司 Cloudera 宣称加大 Spark 框架的投入来取代 MapReduce。

2014 年 5 月，Pivotal Hadoop 集成 Spark 全栈。同月 30 日，Spark 1.0.0 发布。

2015 年，Spark 增加了新的 DataFrames API 和 Datasets API。

2016 年，Spark 2.0 发布。Spark 2.0 与 1.0 的区别主要是，Spark 2.0 修订了 API 的兼容性问题。

2017 年，在美国旧金山举行 Spark Summit 2017，会议介绍 2017 年 Spark 的重点开发方向是深度学习以及对流性能的改进。

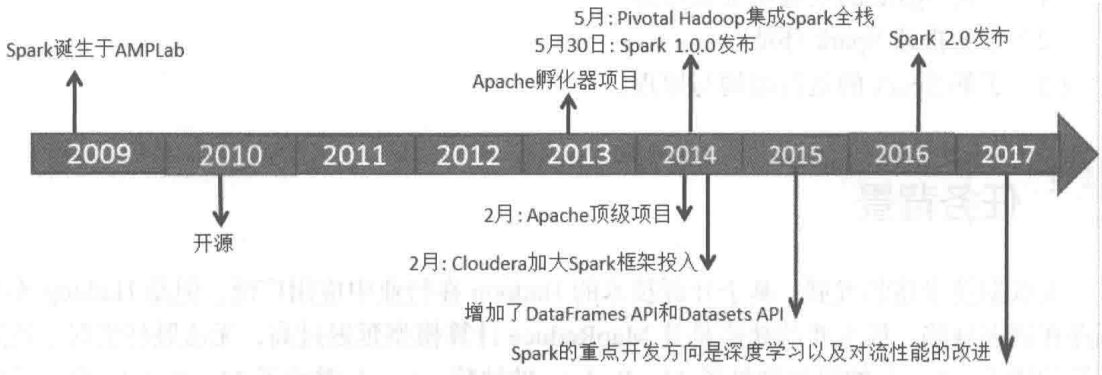


图 1-1　Spark 的发展历史

1.1.2　Spark 的特点

作为新一代轻量级大数据处理平台，Spark 具有以下特点。

1. 快速

图 1-2 所示为分别使用 Hadoop 和 Spark 运行逻辑回归算法，逻辑回归算法一般需要多次迭代。从图中可以看出，Spark 运行逻辑回归算法的速度是 Hadoop MapReduce 运行速度的 100 多倍。一般情况下，对于迭代次数较多的应用程序，Spark 程序在内存中的运行速度是 Hadoop MapReduce 运行速度的 100 多倍，在磁盘上的运行速度是 Hadoop MapReduce 运行速度的 10 多倍。

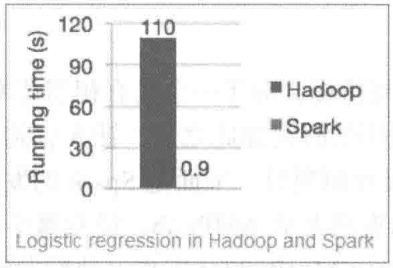


图 1-2　Hadoop MapReduce 与 Spark 运行速度的比较

Spark 与 Hadoop MapReduce 的运行速度为何会有如此大的差异？图 1-3 清晰地解释了两者存在差异的原因。Spark 的中间数据存放于内存中，有更高的迭代运算效率，而 Hadoop

每次迭代的中间数据存放于 HDFS 中，涉及硬盘的读写，明显降低了运算效率。

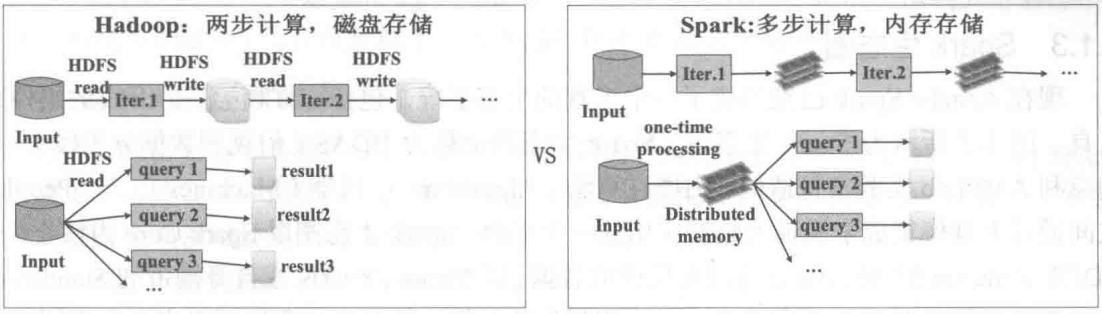


图 1-3 Spark 与 Hadoop 数据存储对比

2. 易用

Spark 支持使用 Scala、Python、Java 及 R 语言快速编写应用。同时 Spark 提供超过 80 个高级运算符，使得编写并行应用程序变得容易，并且可以在 Scala、Python 或 R 的交互模式下使用 Spark。

3. 通用

Spark 可以与 SQL、Streaming 及复杂的分析良好结合。Spark 还有一系列的高级工具，包括 Spark SQL、MLlib（机器学习库）、GraphX（图计算）和 Spark Streaming，并且支持在一个应用中同时使用这些组件，如图 1-4 所示。

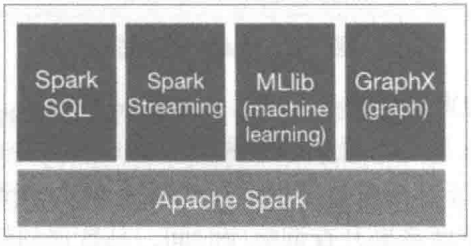


图 1-4 Spark 高级工具架构

4. 随处运行

用户可以使用 Spark 的独立集群模式运行 Spark，也可以在 EC2（亚马逊弹性计算云）、Hadoop YARN 或者 Apache Mesos 上运行 Spark，并且可以从 HDFS、Cassandra、HBase、Hive、Tachyon 和任何分布式文件系统读取数据。

5. 代码简洁

Spark 支持使用 Scala、Python 等语言编写。Scala 或者 Python 的代码相对 Java 来说都比较简洁，因此 Spark 使用 Scala 或者 Python 编写应用程序要比使用 MapReduce 编写应用程序简单方便。比如，MapReduce 实现单词计数可能需要 60 多行代码，而 Spark 用 Scala 语言实现只需要一行，如代码 1-1 所示。

代码 1-1 Spark 实现单词计数代码

```
sc.textFile("/user/root/test.txt").flatMap(_.split("
```

```
")).map((_,1)).reduceByKey(_+_).
saveAsTextFile("/user/root/output")
```

1.1.3 Spark 生态圈

现在 Apache Spark 已经形成了一个丰富的生态系统，包括官方和第三方开发的组件或工具。图 1-5 所示为 Spark 生态圈。Spark 生态圈也称为 BDAS（伯克利数据分析栈），是伯克利 AMPLab 实验室打造的，力图在算法（Algorithms）、机器（Machines）、人（People）之间通过大规模集成来展现大数据应用的一个平台。Spark 生态圈以 Spark Core 为核心，从 HDFS、Amazon S3 和 HBase 等持久层读取数据，以 Mesos、YARN 或自身携带的 Standalone 为资源管理器来调度 Job 完成 Spark 应用程序的计算。这些应用程序可以来自于不同的组件，如 Spark Shell/Spark Submit 的批处理、Spark Streaming 的实时处理应用、Spark SQL 的即席查询、BlinkDB 的权衡查询、MLlib/MLBase 的机器学习、GraphX 的图处理和 SparkR 的数学计算等。

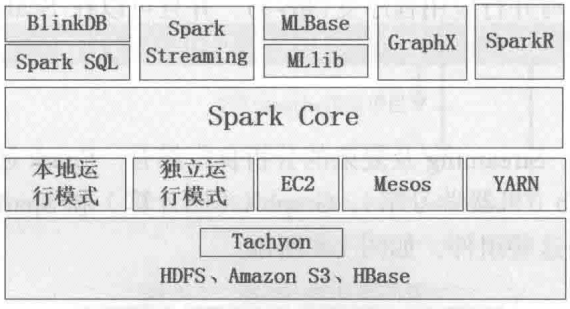


图 1-5 Spark 生态图

下面详细介绍 Spark 的重要组件。

（1）Spark Core：Spark 核心，提供底层框架及核心支持。

（2）BlinkDB：一个用于在海量数据上运行交互式 SQL 查询的大规模并行查询引擎，它允许用户通过权衡数据精度来提升查询响应时间，其数据的精度被控制在允许的误差范围内。

（3）Spark SQL：可以执行 SQL 查询，包括基本的 SQL 语法和 HiveQL 语法。读取的数据源包括 Hive 表、Parquent 文件、JSON 数据、关系数据库（如 MySQL）等。

（4）Spark Streaming：流式计算。比如，一个网站的流量是每时每刻都在发生的，如果需要知道过去 15 分钟或一个小时的流量，则可以使用 Spark Streaming 来解决这个问题。

（5）MLBase：MLBase 是 Spark 生态圈的一部分，专注于机器学习，让机器学习的门槛更低，让一些可能并不了解机器学习的用户也能方便地使用 MLBase。MLBase 分为 4 部分：MLlib、MLI、ML Optimizer 和 MLRuntime。

（6）MLlib：MLBase 的一部分，MLlib 是 Spark 的数据挖掘算法库，实现了一些常见的机器学习算法和实用程序，包括分类、回归、聚类、协同过滤、降维以及底层优化。

（7）GraphX：图计算的应用在很多情况下处理的数据都是很庞大的，比如在移动社交上面的关系等都可以用图相关算法来进行处理和挖掘，但是如果用户要自行编写相关的图计算算法，并且要在集群中应用，那么难度是非常大的。而使用 Spark GraphX 就可以解决

这个问题，它里面内置了很多的图相关算法。

（8）SparkR：SparkR是AMPLab发布的一个R开发包，使得R摆脱单机运行的命运，可以作为Spark的Job运行在集群上，极大地扩展了R的数据处理能力。

1.1.4 Spark的应用场景

目前大数据的应用非常广泛，大数据应用场景的普遍特点是计算量大、效率高。而Spark恰恰满足了这些要求，该项目一经推出便受到开源社区的广泛关注和好评。目前已经发展成为大数据处理领域最炙手可热的开源项目。

以下列举的是国内外应用Spark成功的公司。

1. 腾讯

广点通是最早使用Spark的应用之一。腾讯大数据精准推荐借助Spark快速迭代的优势，围绕“数据+算法+系统”这套技术方案，实现了“数据实时采集、算法实时训练、系统实时预测”的全流程实时并行高维算法，最终成功应用于广点通pCTR投放系统上，支持每天上百亿的请求量。

2. Yahoo

Yahoo将Spark用在Audience Expansion中。Audience Expansion是广告者寻找目标用户的一种方法，首先广告者提供一些观看了广告并且购买产品的样本客户，据此进行学习，寻找更多可能转化的用户，对他们定向广告。Yahoo采用的算法是Logistic Regression。同时由于某些SQL负载需要更高的服务质量，又加入了专门运行Shark的大内存集群，用于取代商业BI/OLAP工具，承担报表/仪表盘和交互式/即席查询，同时与桌面BI工具对接。

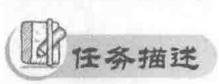
3. 淘宝

淘宝技术团队使用了Spark来解决多次迭代的机器学习算法、高计算复杂度的算法等，将Spark运用于淘宝的推荐相关算法上，同时还利用GraphX解决了许多生产问题，包括以下计算场景：基于度分布的中枢节点发现、基于最大连通图的社区发现、基于三角形计数的关系衡量、基于随机游走的用户属性传播等。

4. 优酷土豆

目前Spark已经广泛使用在优酷土豆的视频推荐、广告业务等方面。Spark交互查询响应快，性能比Hadoop提高了若干倍。一方面，使用Spark模拟广告投放的计算效率高、延迟小（同Hadoop比，延迟至少降低一个数量级）。另一方面，优酷土豆的视频推荐往往涉及机器学习及图计算，而使用Spark解决机器学习、图计算等迭代计算能够大大减少网络传输、数据落地等的次数，极大地提高了计算性能。

任务1.2 搭建Spark环境



Spark环境可分为单机版环境、单机伪分布式环境和完全分布式环境。本节的主要任务是学习如何搭建Spark环境以及查看Spark的服务监控。读者可从官网下载Spark的安装包，本书使用的Spark安装包是spark-1.6.3-bin-hadoop2.6.tgz，下载网址为http://spark.apache.