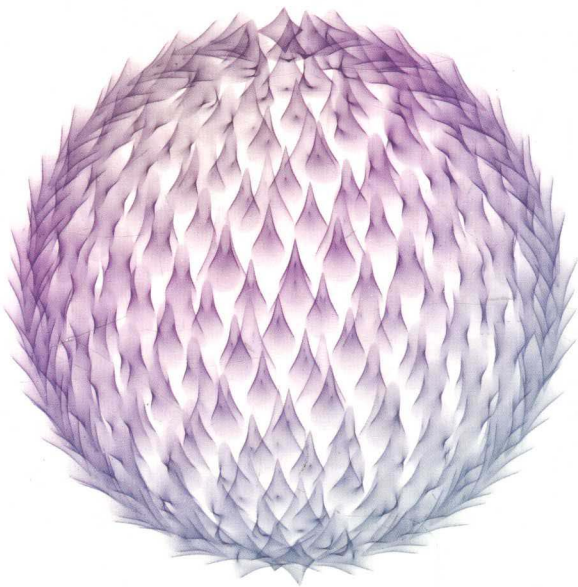
数据中国“百校工程”项目系列教材
数据科学与大数据技术专业系列规划教材

瑞翼教育

Spark
大数据技术与应用

赵红艳 许桂秋 ◉ 主编
潘晓洋 张越 李阳 张军 王露露 ◉ 副主编



BIG DATA
Technology

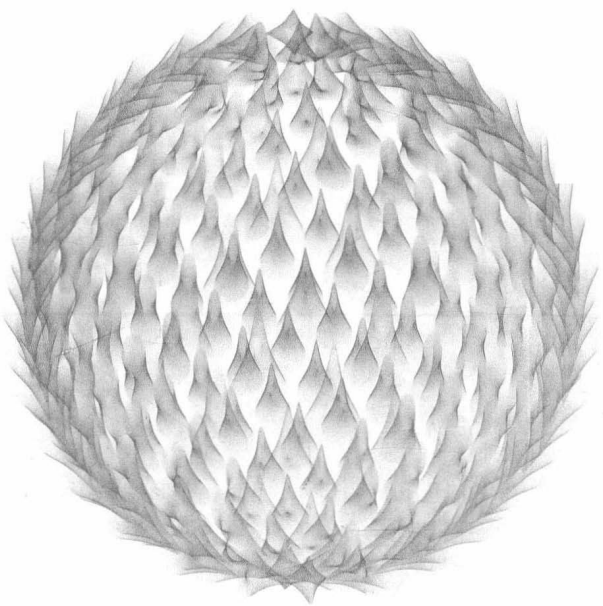
中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

数据中国“百校工程”项目系列教材
数据科学与大数据技术专业系列规划教材

瑞翼教育

Spark
大数据技术与应用

赵红艳 许桂秋◉主编
潘晓洋 张越 李阳 张军 王露露◉副主编



BIG DATA
Technology

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 / 赵红艳，许桂秋主编. -- 北京 : 人民邮电出版社，2019.4
数据科学与大数据技术专业系列规划教材
ISBN 978-7-115-50347-3

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

中国版本图书馆CIP数据核字(2019)第029374号

内 容 提 要

本书采用理论与实践相结合的方式，介绍了 Spark 大数据分析计算框架的基础知识，培养读者使用 Spark 解决实际问题的能力。本书内容包括：Spark 简介与运行原理、Spark 的环境搭建、使用 Python 开发 Spark 应用、Spark RDD、DataFrame 与 Spark SQL、Spark Streaming、Spark 机器学习库、GraphFrames 图计算，并给出了两个综合案例：出租车数据分析、图书推荐系统。

本书可作为高等院校计算机、数据科学与大数据技术等相关专业的教材，也可作为 Spark 开发人员的参考用书。

◆ 主　　编　赵红艳　许桂秋
副 主 编　潘晓洋　张　越　李　阳　张　军　王露露
责任编辑　李　召
责任印制　陈　犇

◆ 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
固安县铭成印刷有限公司印刷

◆ 开本：787×1092　1/16
印张：8.75　　2019 年 4 月第 1 版
字数：214 千字　　2019 年 4 月河北第 1 次印刷

定价：39.80 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

前言

随着信息技术的飞速发展，人类社会正在走向数据技术时代，大数据作为数据技术时代的重要组成部分，背后隐藏的价值逐渐被行业组织、政府部门以及社会公众所认识和关注。与传统的数据相比，大数据具有体量巨大、类型多样、价值密度低、高速性等特点，传统的技术手段已经不能满足大数据处理与分析的要求。

在 Spark 大数据分析计算框架诞生之前，Hadoop 起着重要的作用，在实际的项目中担任数据分析和存储的角色，但在实时性和处理速度等方面均不具备优势，Spark 的出现打破了这种局面。Spark 提供了高级应用程序编程接口，应用开发者只需专注于应用计算本身而不用关注集群。Spark 计算速度快，支持交互式计算和复杂算法。同时，Spark 是一个通用引擎，用户可用它来完成各种运算，包括 SQL 查询、文本处理、机器学习等。

本书作为大数据分析与内存计算的入门教材，希望能够帮助读者打开大数据分析与应用的大门。全书采用理论结合实践的方式，循序渐进地介绍了 Spark 的运行原理，并引入综合性的实践案例，引导读者运用所学知识解决现实中的问题。

全书共分为 10 章。第 1 章主要介绍了 Spark 的发展历程、起源和基本的运行原理，为后面章节学习 Spark 组件打下基础。第 2 章介绍了 Spark 的环境搭建，从软件下载、解压到环境变量、相应参数的设置等方面讲解了如何在本地搭建 Spark 环境。第 3 章主要介绍了 Python 编程语言、PySpark 的启动和日志的设置、PySpark 开发包的安装以及如何使用开发工具 PyCharm 编写 Spark 应用。第 4 章主要介绍了 Spark RDD 弹性分布式数据集，内容包括 RDD 的特点、创建 RDD 的过程、RDD 的常用算子以及 RDD 的持久化。第 5 章主要介绍了 DataFrame 与 Spark SQL 的区别与联系，并对常用操作进行了阐释。第 6 章介绍了 Spark Streaming 的特点、不同数据源的加载、DStream 输出与转换操作以及与 DataFrame 和 SQL 的互操作。第 7 章主要介绍了机器学习的相关概念、机器学习的一般流程、Spark 机器学习库 MLlib 和 ML。第 8 章主要介绍了基于 Spark 的第三方开源图计算工具 GraphFrames，内容包括图的概念、图的操作和图算法，并用实际案例进行了应用操作。第 9 章和第 10 章分别介绍了综合实例：出租车数据分析和图书推荐系统，通过两个综合实例让读者从理论阶段上升到实际应用阶段，帮助读者更深入理解 Spark 的使用方法。

本课程建议安排 64 课时，教师可根据学生的实际学习情况以及高校的培养方案选择教学内容。本书可以作为高等院校计算机、数据科学与大数据技术等相关专业的教材，也可作为 Spark 开发人员的参考用书。

由于编者水平有限，书中难免出现一些疏漏和不足，恳请广大读者批评指正。

编者

2019 年 1 月

目 录

第 1 章 Spark 简介与运行原理

Spark 是现在流行的大数据分析计算框架，在大数据应用中起着不可或缺的作用。本章从 Spark 的产生、发展及其生态圈等方面对 Spark 进行介绍。

本章主要内容如下。

（1）Spark 是什么。

（2）Spark 的生态系统。

（3）Spark 架构与原理。

（4）Spark 2.X 新特性。

1.1 Spark 是什么

Spark 是 2009 年由马泰·扎哈里亚（Matei Zaharia）在加州大学伯克利分校的 AMPLab 实验室开发的子项目，经过开源后捐赠给 Apache 软件基金会，最后成为我们现在众所周知的 Apache Spark。它是由 Scala 语言实现的专门为大规模数据处理而设计的快速通用的计算引擎。经过多年的发展，现已形成了一个高速发展、应用广泛的生态系统。

Spark 主要有以下 3 个特点。

（1）Spark 提供了高级应用程序编程接口（Application Programming Interface，API），应用开发者只用专注于应用计算本身即可，而不用关注集群。

（2）Spark 计算速度快，支持交互式计算和复杂算法。

（3）Spark 是一个通用引擎，可用它来完成各种运算，包括 SQL 查询、文本处理、机器学习、实时流处理等。在 Spark 出现之前，我们一般需要学习使用各种各样的大数据分析引擎来分别实现这些需求。

1.1.1 Spark 的版本发展历程

Spark 从诞生至今迭代了很多个版本，其性能和生态也是越来越好，目前已经升级到 2.3.2 版本。其主要发展历程如表 1-1 所示。

表 1-1　　Spark 的版本发展历程

年代	说明
2009	Spark 由 Matei Zaharia 在加州大学伯克利分校的 AMPLab 实验室开发
2010	通过 BSD 授权条款发布开放源码
2013	Spark 项目被捐赠给 Apache 软件基金会
2014/2	Spark 成为 Apache 的顶级项目
2014/11	Databricks 团队使用 Spark 刷新数据排序的世界纪录
2015/3	Spark 1.3.0 版本发布，开始加入 DataFrame 与 SparkML
2016/7	Spark 2.0.0 版本发布，提升执行性能，更容易被使用
2017/7	Spark 2.2.0 版本发布，从结构化流中删除实验标签
2018/2	Spark 2.3.0 版本发布，增加对结构流连续处理的支持
2018/9	Spark 2.3.2 版本发布

1.1.2 Spark 与 Hadoop 的区别与联系

Spark 与 Hadoop 处理的许多任务相同，但是在以下两个方面不相同。

（1）解决问题的方式不一样

Hadoop 和 Spark 两者都是大数据框架，但是各自的属性和性能却不完全相同。Hadoop 是一个分布式数据基础设施，它将巨大的数据集分派到一个由普通计算机组成的集群中的多个节点进行存储，这意味着我们不需要购买和维护昂贵的服务器硬件。同时，Hadoop 还会对这些数据进行排序和追踪，这使得大数据处理和分析更加迅速高效。

Spark 则是一个专门用来对分布式存储的大数据进行处理的工具，但它并不会进行分布式数据的存储。

（2）两者可合可分

Hadoop 不仅提供了 HDFS 分布式数据存储功能，还提供了 MapReduce 的数据处理功能。因此我们可以不使用 Spark，而选择使用 Hadoop 自身的 MapReduce 对数据进行处理。

同样，Spark 也不一定需要依附在 Hadoop 系统中。但如上所述，因为 Spark 没有提供文件管理系统，所以它需要和其他的分布式文件系统先进行集成然后才能运作。

1.1.3　Spark 的应用场景

Spark 使用了内存分布式数据集技术，除了能够提供交互式查询外，它还提升了迭代工作负载的性能。在互联网领域，Spark 有快速查询、实时日志采集处理、业务推荐、定制广告、用户图计算等强大功能。国内外的一些大公司，比如谷歌（Google）、阿里巴巴、英特尔（Intel）、网易、科大讯飞等都有实际业务运行在 Spark 平台上。

下面简单介绍一下 Spark 在各个领域中的用途。

（1）快速查询系统。基于日志数据的快速查询系统业务构建于 Spark 之上，利用其快速查询和内存表等优势，Spark 能够承担大多数日志数据的即时查询工作，在性能方面普遍比 Hive 快 2～10 倍。如果借助内存表的功能，性能将会比 Hive 快百倍。

（2）实时日志采集处理系统。Spark 流处理模块对业务日志进行实时快速迭代处理，并进行综合分析，用来满足线上系统分析要求。

（3）业务推荐系统。Spark 将业务推荐系统的小时和天级别的模型训练，转变为分钟级别的模型训练；能有效地优化相关排名、个性化推荐以及热点分析等。

（4）定制广告系统。定制广告业务需要大数据做应用分析、效果分析、定向优化等，借助 Spark 快速迭代的优势，可以实现在“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，支持上亿的请求量处理。模拟广告投放计算延迟小、效率高，同 MapReduce 相比，延迟至少降低一个数量级。

（5）用户图计算。利用 Spark 图计算解决了许多生产问题，如基于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。

1.2　Spark 的生态系统

Spark 生态系统以 Spark Core 为核心，利用 Standalone、YARN 和 Mesos 等进行资源调度管理，完成应用程序分析与处理。这些应用程序来自 Spark 的不同组件，如 Spark Shell、Spark Submit 交互式批处理、Spark Streaming 实时流处理、Spark SQL 快速查询、MLlib 机器学习、GraphX 图处理等，如图 1-1 所示。

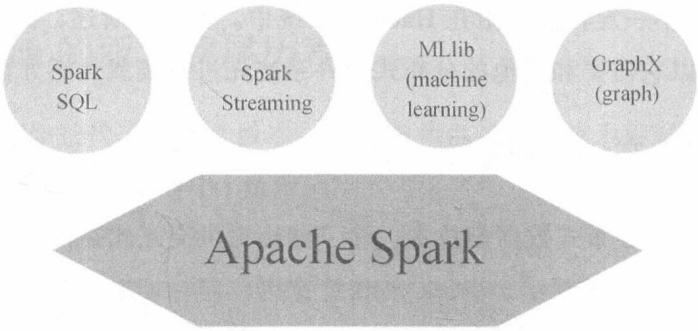


图 1-1　Spark 生态系统图

（1）Spark Core 提供 Spark 最基础与最核心的功能，它的子框架包括 Spark SQL、Spark Streaming、MLlib 和 GraphX。

（2）Spark Streaming 是 Spark API 核心的一个存在可达到超高通量的扩展，可以处理实时数据流的数据并进行容错。它可以从 Kafka、Flume、Twitter、ZeroMQ、Kinesis、TCP sockets 等数据源获取数据，并且可以使用复杂的算法和高级功能对数据进行处理。处理后的数据可以被推送到文件系统或数据库。

（3）Spark SQL 是一种结构化的数据处理模块。它提供了一个称为 Data Frame 的编程抽象，也可以作为分布式 SQL 查询引擎。

一个 DataFrame 相当于一个列数据的分布式采集组织，类似于一个关系型数据库中的一个表。它可以从多种方式构建，如结构化数据文件、Hive、外部数据库或分布式动态数据集（RDD）。

（4）GraphX 在 Graphs 和 Graph-parallel 并行计算中是一个新的部分，GraphX 是 Spark 上的分布式图形处理架构，可用于图表计算。

1.3　Spark 的架构与原理

1.3.1　Spark 架构设计

Spark 架构主要包括客户端驱动程序 Driver App、集群管理器 Cluster Manager、工作节点 Worker 以及基本任务执行单元 Executor。

（1）Driver App 是客户端驱动程序，也可以理解为客户端应用程序，用于将任务程序转换为 RDD 和 DAG，并与 Cluster Manager 进行通信与调度。

（2）Cluster Manager 是 Spark 的集群管理器。它主要负责资源的分配与管理。集群管

理器分配的资源属于一级分配，它将各个 Worker 上的内存、CPU 等资源分配给应用程序，但是并不分配 Executor 的资源。目前，Standalone、YARN、Mesos、EC2 等都可以作为 Spark 的集群管理器。

（3）Worker 是 Spark 的工作节点。对 Spark 应用程序来说，由集群管理器分配得到资源的 Worker 节点主要负责以下工作：创建 Executor，将资源和任务进一步分配给 Executor，然后同步资源信息给 Cluster Manager。

（4）Executor 是 Spark 任务的执行单元。它主要负责任务的执行以及与 Worker、Driver App 的信息同步。

Spark 架构设计图如图 1-2 所示。

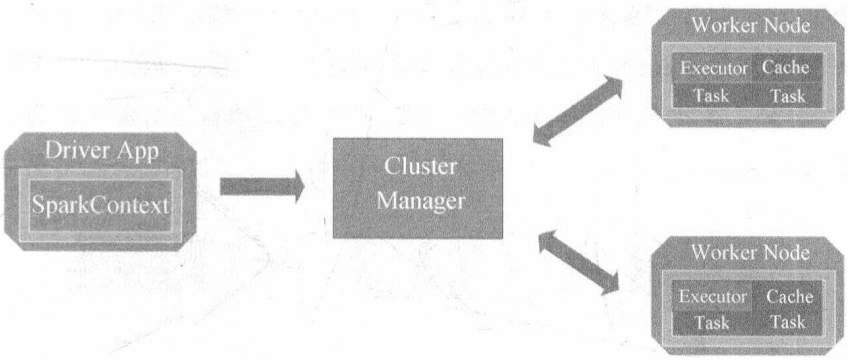


图 1-2　Spark 架构设计图

1.3.2　Spark 作业运行流程

Spark 作业流程图如图 1-3 所示。

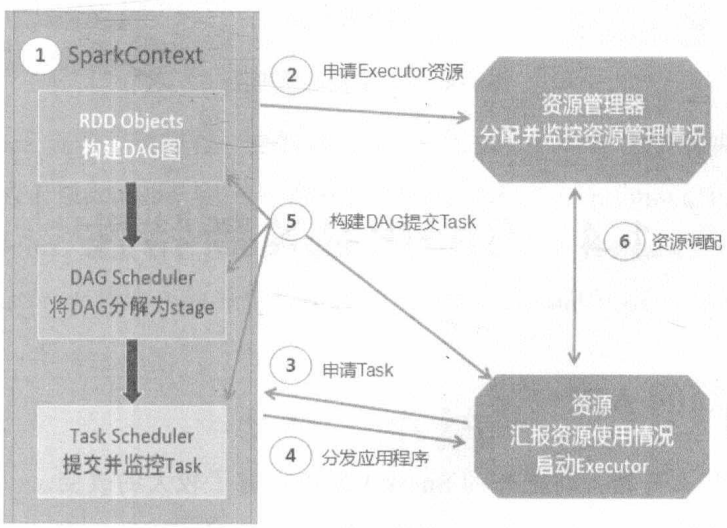


图 1-3　Spark 作业流程图

（1）构建 Spark Application 的运行环境，启动 SparkContext。

（2）SparkContext 向资源管理器申请运行 Executor 资源，并启动 StandaloneExecutorbackend。

（3）Executor 向 SparkContext 申请 Task。

（4）SparkContext 将应用程序分发给 Executor。

（5）SparkContext 构建 DAG 图，将 DAG 图分解成 Stage，将 Taskset 发送给 Task Scheduler，由 Task Scheduler 将 Task 发送给 Executor 运行。

（6）Task 在 Executor 上运行，运行完释放所有资源。

1.3.3 Spark 分布式计算流程

Spark 分布式计算流程包含以下几个步骤：首先分析应用的代码创建有向无环图，然后将有向无环图划分为 Stage，然后 Stage 生成作业（job），生成作业后由 FinalStage 提交任务集，提交任务的工作交给 TaskSets 完成，然后每个 Task 执行所分配的任务，最终 Results 跟踪结果。流程图如图 1-4 所示。

图 1-4　Spark 核心原理图

1.4　Spark 2.X 新特性

1.4.1 精简的 API

从 Spark 2.0 版本开始，与之前的 Spark 1.X 版本有了较大的改变。

（1）统一的 DataFrame 和 Dataset 接口。统一了 Scala 和 Java 的 DataFrame、Dataset

接口，在 R 和 Python 中由于缺乏安全类型，DataFrame 成为主要的程序接口。

（2）新增 SparkSession 入口。SparkSession 替代原来的 SQLContext 和 HiveContext 作为 DataFrame 和 Dataset 的入口函数。SQLContext 和 HiveContext 保持向后兼容。

（3）为 SparkSession 提供全新的工作流式配置。

（4）更易用、更高效的计算接口。

（5）Dataset 中的聚合操作有全新的、改进的聚合接口。

1.4.2　Spark 作为编译器

Spark 2.0 搭载了第二代 Tungsten 引擎，该引擎是根据现代编译器与 MPP 数据库的理念来构建的，它将这些理念用于数据处理中，其主要思想就是在运行时使用优化后的字节码，将整体查询合成为单个函数，不再使用虚拟函数调用，而是利用 CPU 来注册中间数据。

为了有直观的感受，表 1-2 显示了 Spark 1.6 与 Spark 2.0 分别在一个核上处理一行的操作时间（单位：ns）。

表 1-2　　Spark 1.6 与 Spark 2.0 操作时间对比图　　单位：ns

原生的函数	Spark 1.6	Spark 2.0
filter	15	1.1
sum w/o group	14	0.9
sum w/ group	79	10.7
hash join	115	4.0
sort (8-bit entropy)	620	5.3
sort (64-bit entropy)	620	40
sort-merge join	750	700

1.4.3　智能化程度

为了实现 Spark 更快、更轻松、更智能的目标，Spark 2.X 在许多模块上都做了重要的更新，如 Structured Streaming 引入了低延迟的连续处理（Continuous Processing）、支持 Stream-to-stream Joins、通过改善 Pandas UDFs 的性能来提升 PySpark、支持第 4 种调度引擎 Kubernetes Clusters（其他 3 种分别是自带的独立模式 Standalone、YARN、Mesos）等。

1.5　小结

本章主要介绍了 Spark 定义、生态系统、架构原理和新特性等内容，从原理到应用由

浅入深地介绍了 Spark，让读者从宏观和微观上对 Spark 有了认识和了解。

习　　题

简答题

（1）Spark 与 Hadoop 的区别是什么？

（2）Spark 的应用场景有哪些？

（3）简述 Spark 的作业运行流程。

（4）Spark 2.X 与 Spark 1.X 有什么不同点？