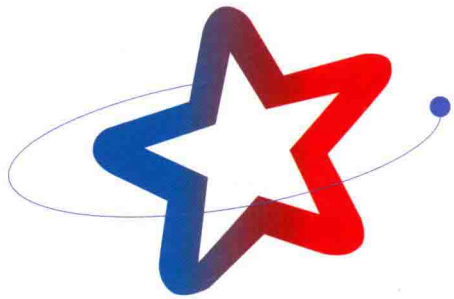
- 讲解Spark SQL背景知识、编程基础
- 通过一个工程实例让读者领略Spark SQL强大简便的分析能力
- 学习构建一个大数据实际应用的方法，加强工程思维
- 洞悉Spark的调优方式及其思想，让Spark SQL程序高效运行

大数据 技术

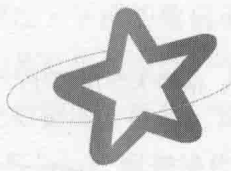


Lightning-fast cluster computing

Spark SQL 入门与实践指南

纪 涵 靖晓文 赵政达 著

清华大学出版社



Spark SQL 入门与实践指南

纪 涵 靖晓文 赵政达 著

清華大學出版社
北京

内 容 简 介

Spark SQL 是 Spark 大数据框架的一部分，支持使用标准 SQL 查询和 HiveQL 来读写数据，可用于结构化数据处理，并可以执行类似 SQL 的 Spark 数据查询，有助于开发人员更快地创建和运行 Spark 程序。

全书分为 4 篇，共 9 章，第一篇讲解了 Spark SQL 发展历史和开发环境搭建。第二篇讲解了 Spark SQL 实例，使得读者掌握 Spark SQL 的入门操作，了解 Spark RDD、DataFrame 和 DataSet，并熟悉 DataFrame 各种操作。第三篇讲解了基于 WiFi 探针的商业大数据分析项目，实例中包含数据采集、预处理、存储、利用 Spark SQL 挖掘数据，一步一步带领读者学习 Spark SQL 强大的数据挖掘功能。第四篇讲解了 Spark SQL 优化的知识。

本书适合 Spark 初学者、Spark 数据分析人员以及 Spark 程序开发人员，也适合高校和培训学校相关专业的师生教学参考。

图书在版编目（CIP）数据

Spark SQL 入门与实践指南 / 纪涵，靖晓文，赵政达著. — 北京：清华大学出版社，2018

ISBN 978-7-302-49670-0

I. ①S… II. ①纪… ②靖… ③赵… III. ①数据处理软件－指南 IV. ①TP274-62

中国版本图书馆 CIP 数据核字（2018）第 034811 号

责任编辑： 夏毓彦
封面设计： 王　翔
责任校对： 闫秀华
责任印制： 杨　艳

出版发行： 清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　邮　　编：100084
社 总 机：010-62770175　　邮　　购：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn

印 刷 者： 北京富博印刷有限公司
装 订 者： 北京市密云县京文制本装订厂
经　　销： 全国新华书店
开　　本： 190mm×260mm　　**印　　张：** 13.25　　**字　　数：** 339 千字
版　　次： 2018 年 4 月第 1 版　　**印　　次：** 2018 年 4 月第 1 次印刷
印　　数： 1~3500
定　　价： 49.00 元

产品编号：076458-01

前 言

我们处于一个数据爆炸的时代！

大量涌现的智能手机、平板、可穿戴设备及物联网设备每时每刻都在产生新的数据，然而带来革命性变革的并非海量数据本身，而是我们如何从这些数据中挖掘到有价值的信息，来辅助我们做出更加智能的决策。我们知道，在生产环境下，所谓的大数据往往是由数千万条、上亿条具有多个预定义字段的数据单元组成的数据集，是不是很像传统关系型数据库的二维数据表呢？那么我们是否也能找到一个像 SQL 查询那样简便的工具来高效地分析处理大数据领域中的海量结构化数据呢？没错，这个工具就是 Spark SQL。

Spark SQL 是 Spark 用来操作结构化数据的高级模块，在程序中通过引入 Spark SQL 模块，我们便可以像从前在关系型数据库利用 SQL（结构化查询语言）分析关系型数据库表一样简单快捷地在 Spark 大数据分析平台上对海量结构化数据进行快速分析，而 Spark 平台屏蔽了底层分布式存储、计算、通信的细节以及作业解析、调度的细节，使我们开发者仅需关注如何利用 SQL 进行数据分析的程序逻辑就可以方便地操控集群来分析我们的数据。

本书内容

本书共分为四篇：入门篇、基础篇、实践篇、调优篇，所有代码均采用简洁而优雅的 Scala 语言编写，Spark 框架也是使用 Scala 语言编写的。

第一部分　入门篇（第 1、2 章）

第 1 章简要介绍 Spark 的诞生、Spark SQL 的发展历史以及 Spark SQL 的用处等内容，使读者快速了解 Spark SQL 背景知识，为以后的学习奠定基础。

第 2 章通过讲解 Spark SQL 开发环境的搭建、Spark 作业的打包提交、常见问题的解答，并结合大量图示，使读者快速掌握开发环境的搭建以及提交应用程序到集群上，为后面章节的学习奠定坚实的基础。

第二部分　基础篇（第 3、4、5、6 章）

第 3 章是真正开始学习 Spark SQL 必要的先修课，其中详尽地介绍了 Spark 框架对数据的核心抽象——RDD（弹性分布式数据集）的方方面面。先介绍与 RDD 相关的基本概念，例如

转化操作、行动操作、惰性求值、缓存，讲解的过程伴随着丰富的示例，旨在提高读者对 RDD 的理解与加强读者的 RDD 编程基础。在讲明白 RDD 中基础内容的同时，又深入地剖析了疑点、难点，例如 RDD Lineage（RDD 依赖关系图）、向 Spark 传递函数、对闭包的理解等。在之前对基本类型 RDD 的学习基础上，又引入了对特殊类 RDD——键值对 RDD 的大致介绍，在键值对 RDD 介绍中对 combineByKey 操作的讲解，深入地从代码实现的角度洞悉了 Spark 分布式计算的实质，旨在帮助对 RDD 有着浓厚兴趣的读者做进一步的拓展。最后，站在 RDD 设计者的角度重新审视了 RDD 缓存、持久化、checkpoint 机制，从而诠释了 RDD 为什么能够很好地适应大数据分析业务的特点，有天然强大的容错性、易恢复性和高效性。

第 4 章对 Spark 高级模块——Spark SQL，也就是本书的主题，进行了简明扼要的概述，并讲述了相应的 Spark SQL 编程基础。先是通过与前一章所学的 Spark 对数据的核心抽象——RDD 的对比，引出了 Spark SQL 中核心的数据抽象——DataFrame，讲解了两者的异同，点明了 Spark SQL 是针对结构化数据处理的高级模块的原因在于其内置丰富结构信息的数据抽象。后一部分通过丰富的示例讲解了如何利用 Spark SQL 模块来编程的主要步骤，例如，从结构化数据源中创建 DataFrames、DataFrames 基本操作以及执行 SQL 查询等。

第 5、6 章属于 Spark SQL 编程的进阶内容，也是我们将 Spark SQL 应用于生产、科研计算环境下，真正开始分析多类数据源、实现各种复杂业务需求必须要掌握的知识。在第 5 章里，我们以包含简单且典型的学生信息表的 JSON 文件作为数据源，深入对 DataFrame 丰富强大的 API 进行研究，以操作讲解加示例的形式包揽了 DataFrame 中每一个常用的行动、转化操作，进而帮助读者轻松高效地组合使用 DataFrame 所提供的 API 来实现业务需求。在第 6 章里，介绍了 Spark SQL 可处理的各种数据源，包括 Hive 表、JSON 和 Parquet 文件等，从广度上使读者了解 Spark SQL 在大数据领域对典型结构化数据源的皆可处理性，从而使读者真正在工作中掌握一门结构化数据的分析利器。

第三部分　实践篇（第 7、8 章）

第 7 章通过讲解大型商业实例项目（基于 WiFi 探针的商业大数据分析技术）的功能需求、系统架构、功能设计、数据库结构来帮助读者理解如何在实际开发中应用 Spark SQL 来处理结构化数据，加强读者的工程思维，同时为第 8 章的学习做好铺垫。

第 8 章通过讲解分布式环境搭建以及项目代码的解析来帮助读者进一步理解 Spark SQL 应用程序的执行过程，在后一部分介绍了 Spark SQL 程序的远程调试方法和 Spark 的 Web 界面，帮助读者更加方便地了解程序的运行状态。

第四部分　调优篇（第 9 章）

调优篇由第 9 章组成，本篇从 Spark 的执行流程到内存以及任务的划分，再到 Spark 应用程序的编写技巧，接着到 Spark 本身的调优，最后引出数据倾斜的解决思路，层层递进，逐步解析 Spark 的调优思想。最后以对 Spark 执行引擎 Tungsten 与 Spark SQL 的解析引擎 Catalyst

的介绍作为本部分的结尾。笔者将在本篇中带领读者掌握 Spark 的调优方式以及思想，让 Spark 程序再快一点。

本书适合读者

本书适合于学习数据挖掘、有海量结构化数据分析需求的大数据从业者及爱好者阅读，也可以作为高等院校相关专业的教材。建议在学习本书内容的过程中，理论联系实际，独立进行一些代码的编写，采取开放式的实验方法，即读者自行准备实验数据和实验环境，解决实际问题，最终达到理论联系实际的目的。

本书在写作过程中得到了家人以及本书编辑的大力支持，在此对他们一并表示感谢。

本书由纪涵（主要负责基础篇的编写）主笔，其他参与著作的还有靖晓文（主要负责实践篇的编写）、赵政达（主要负责入门篇、调优篇的编写），排名不分先后。

纪 涵

2018 年 2 月

目 录

第四部分 优化篇

第一部分　入门篇

本书的第一部分由第 1 章和第 2 章组成。第 1 章主要从 Spark SQL 的由来以及 Spark SQL 能做什么两方面对 Spark SQL 进行简单的介绍。第 2 章介绍 Spark 程序编写环境的搭建和 Spark 程序的打包及提交。

第一部分 入门篇

第 1 章
初识Spark SQL

在这一章中读者将大致了解 Spark SQL 的发展历程、Spark SQL 的特点，以及用 Spark SQL 能做些什么。

1.1 Spark SQL 的前世今生

1. Spark 的诞生

相信大家都听说过 MapReduce 这个框架，MapReduce 是对分布式计算的一种抽象。程序员对 map 方法和 reduce 方法进行简单的编写就能迅速地构建出并行化的程序，而不用担心工作在集群上的分布和集群当中数据的容错，这就大大地降低了程序的编写和部署难度。

遗憾的是，MapReduce 这个框架缺少了对分布式内存利用的抽象，这就导致了在不同的计算任务间（比如说两个 MapReduce 工作之间）对数据重用的时候只能采用将数据写回到硬盘中的方法。而计算机将数据写回到磁盘的这个过程耗时是很长的。如今，许多机器学习的算法都需要对数据进行重用，并且这些算法中都包含着大量的迭代计算，比如说 PageRank、K-means 等算法。如果使用 MapReduce 来实现这些算法，那么在执行的时候，将会有大量的时间被消耗在 I/O 上面。针对这个问题，伯克利大学提出了 RDDs（弹性分布式数据集 RDDs 是一个具有容错性和并行性的数据结构，它可以让我们将中间结果持久化到内存中）的思想，RDDs 提供了对内存的抽象，然后伯克利大学根据 RDDs 的思想设计出了一个系统，Spark 就这样诞生了。

2. 从 Shark 到 Spark SQL

Spark 诞生之后，人们开始使用 Spark，并且喜欢上了 Spark。渐渐的，使用 Spark 的人越来越多。突然有一天，一部分人产生了一个大胆的想法：Hadoop 上面有 Hive，Hive 能把 SQL 转成 MapReduce 作业，这多么方便啊！Spark 这么好用的系统却没有配备类似 Hive 这样的工具，要不我们也造一个这样的工具吧！于是 Shark 被提了出来，Shark 将 SQL 语句转成 RDD 执行。这就仿照了 Hadoop 生态圈，做出了一个 Spark 版本的“Hive”。做出这个工具之后人们十分开心，因为他们终于也能愉快地使用 SQL 对数据进行查询分析了，可以大大地提高程序的编写效率。图 1-1 所示是 Shark 的架构示意图，来自 https://amplab.cs.berkeley.edu/wp-content/

uploads/2012/03/mod482-xin1.pdf。

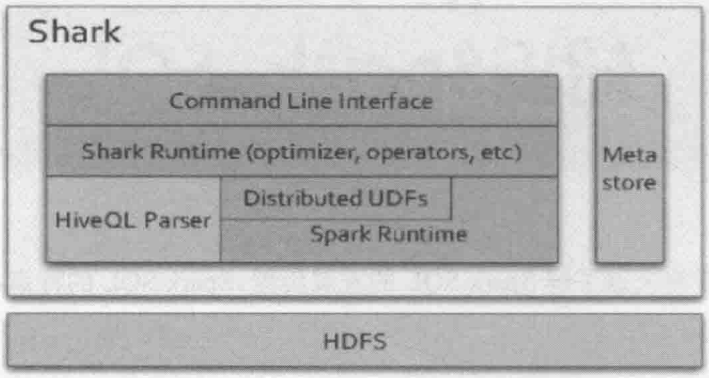


图 1-1

随着越来越多的人使用 Shark 和其版本的更新，人们发现 Shark 具有一定的局限性。细心的小伙伴会发现图 1-1 中 Shark 的框架使用了 HiveQL Parser 这一模块。这样一来 Shark 对 Hive 有了依赖，导致 Shark 添加一些新的功能或者修改一些东西时特别不方便。这样 Shark 的发展受到了严重的限制。

由于 Shark 这样的一些弊端，在 2014 年左右人们决定终止 Shark 这个项目并且将精力转移到 Spark SQL 的研发当中去。之后一个新的 SQL 引擎——Spark SQL 就诞生了。

1.2 Spark SQL 能做什么

现在我们知道了 Spark SQL 是怎么来的，那么 Spark SQL 到底能做些什么呢？下面我们根据 ETL（数据的抽取、转换、加载）的三个过程来讲解一下 Spark SQL 的作用。

（1）抽取（Extract）：Spark SQL 可以从多种文件系统（HDFS、S3. 本地文件系统等）、关系型数据库（MySQL、Oracle、PostgreSQL 等）或 NoSQL 数据库（Cassandra、HBase、Druid 等）中获取数据，Spark SQL 支持的文件类型可以是 CSV、JSON、XML、Parquet、ORC、Avro 等。得益于 Spark SQL 对多种数据源的支持，Spark SQL 能从多种渠道抽取人们想要的数据到 Spark 中。

（2）转换（Transform）：我们常说的数据清洗，比如空值处理、拆分数据、规范化数据格式、数据替换等操作。Spark SQL 能高效地完成这类转换操作。

（3）加载（Load）：在数据处理完成之后，Spark SQL 还可以将数据存储到各种数据源（前文提到的数据源）中。

如果你以为 Spark SQL 只能做上面这些事情，那你就错了。Spark SQL 还可以作为一个分布式 SQL 查询引擎通过 JDBC 或 ODBC 或者命令行的方式对数据库进行分布式查询。Spark

SQL 中还有一个自带的 Thrift JDBC/ODBC 服务，可以用 Spark 根目录下的 sbin 文件夹中的 start-thriftserver.sh 脚本启动这个服务。Spark 中还自带了一个 Beeline 的命令行客户端，读者可以通过这个客户端连接启动的 Thrift JDBC/ODBC，然后提交 SQL。

如果你以为 Spark SQL 能做的只有这些，那你就错了。Spark SQL 还可以和 Spark 的其他模块搭配使用，完成各种各样复杂的工作。比如和 Streaming 搭配处理实时的数据流，和 MLlib 搭配完成一些机器学习的应用。

第 2 章
Spark安装、编程环境搭建以及打包提交

通过上一章的学习，相信读者已经了解了 Spark SQL 是什么、能做什么、发展状况如何，在这一章中读者将学习在 Linux 中完成 Spark 的安装，以及搭建本书后面需要用到的 Spark 程序的编写环境，并能够将程序打包提交到 Spark 中运行。

2.1 Spark 的简易安装

搭建 Spark 之前需要读者先安装好 Hadoop，由于这个环境用于本书学习，这里建议部署单机或者伪分布式的 Hadoop。另外，关于 Hadoop 的安装这里不予以介绍，大家可自行搜集 Hadoop 安装教程，确保 HDFS 能正常使用即可。Spark 2.2.0 官网中明确表明了：Spark 2.2.0 不支持 Java 7、Python 2.6 以及 Hadoop 2.6.5 之前的版本。笔者使用的系统是 CentOS 7、Java 8、Hadoop 2.7.3，这里配的 Spark 是单机模式。

步骤 01 下载 Spark 安装包。

进入 Spark 的下载页面 https://spark.apache.org/downloads.html，如图 2-1 所示。

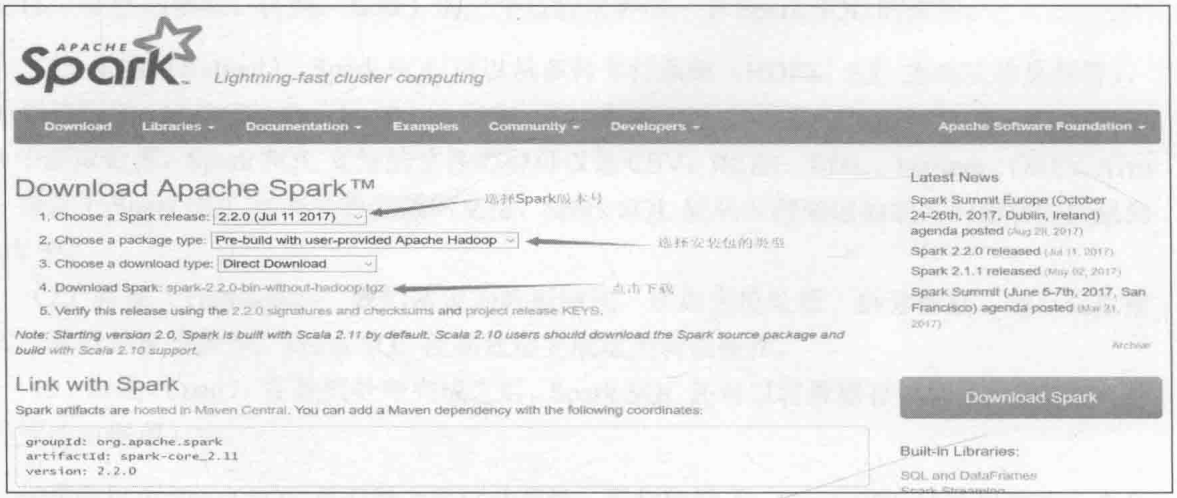


图 2-1