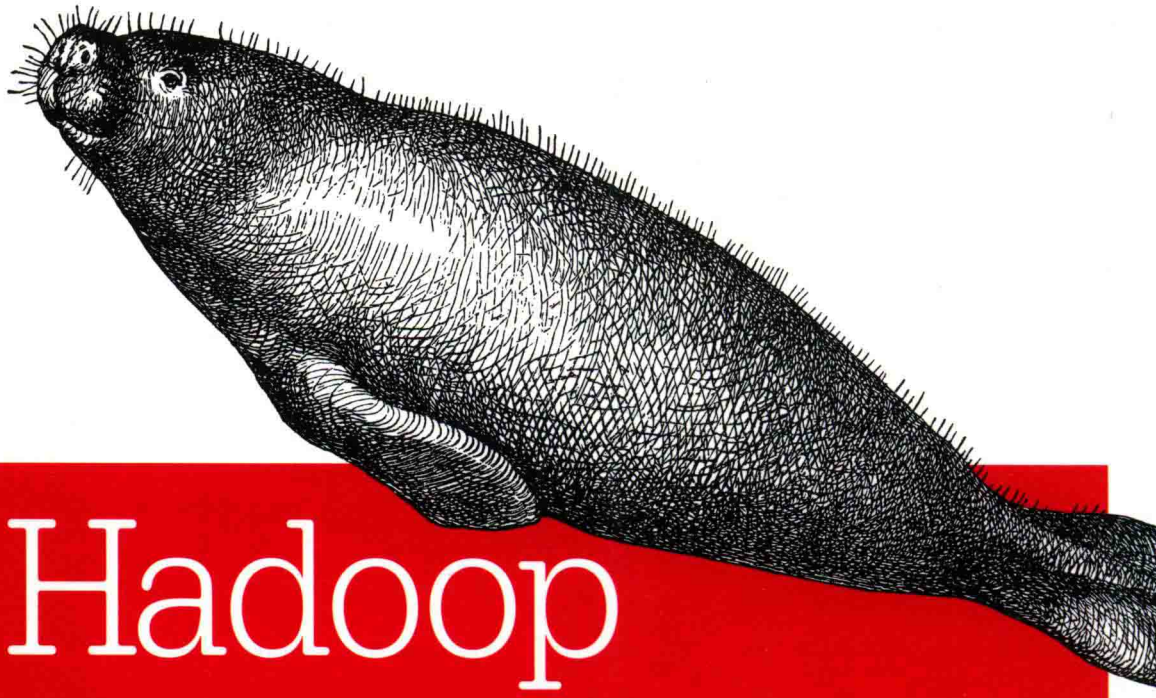
O'REILLY®

TURING 图灵程序设计丛书



Hadoop 应用架构

Hadoop Application Architectures

偏重Hadoop实践，直击企业大数据管理痛点，全面解析应用构架
阐述如何有效集成MapReduce、Spark、Hive等工具以形成完整数据解决方案

[美] Mark Grover, Ted Malaska,
Jonathan Seidman, Gwen Shapira 著
郭文超 译

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

TURING 图灵程序设计丛书

Hadoop应用架构

Hadoop Application Architectures

[美] Mark Grover, Ted Malaska, Jonathan Seidman, Gwen Shapira 著
郭文超 译

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'Reilly Media, Inc.授权人民邮电出版社出版

人民邮电出版社
北 京

图书在版编目（CIP）数据

Hadoop应用架构 / （美）马克·格洛沃（Mark Grover）等著 ; 郭文超译. -- 北京 : 人民邮电出版社, 2017.1
（图灵程序设计丛书）
ISBN 978-7-115-44243-7

Ⅰ. ①H… Ⅱ. ①马… ②郭… Ⅲ. ①数据处理软件 Ⅳ. ①TP274

中国版本图书馆CIP数据核字(2016)第292044号

内 容 提 要

本书讲解使用Hadoop平台进行应用架构所需要的关键知识，旨在帮助读者掌握有效集成HBase、Kafka、Spark等Hadoop生态圈工具以形成完整的大数据解决方案。书中内容分为两部分，第一部分介绍使用Hadoop创建应用程序时要考虑的问题，第二部分展示如何使用前面介绍的组件实现基于Hadoop的完整解决方案。

本书适合软件开发人员、构架师、项目主管等。

◆ 著　[美] Mark Grover，Ted Malaska，Jonathan Seidman，Gwen Shapira
译　郭文超
责任编辑　朱　巍
执行编辑　张　慜
责任印制　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
北京昌平百善印刷厂印刷

◆ 开本：800×1000　1/16
印张：19
字数：449千字　2017年1月第1版
印数：1－4 000册　2017年1月北京第1次印刷
著作权合同登记号　图字：01-2016-7590号

定价：69.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广字第8052号

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

站在巨人的肩上

Standing on Shoulders of Giants

iTuring.cn

版权声明

O'Reilly Media, Inc.介绍

O'Reilly Media通过图书、杂志、在线服务、调查研究和会议等方式传播创新知识。自1978年开始，O'Reilly一直都是前沿发展的见证者和推动者。超级极客们正在开创着未来，而我们关注真正重要的技术趋势——通过放大那些“细微的信号”来刺激社会对新科技的应用。作为技术社区中活跃的参与者，O'Reilly的发展充满了对创新的倡导、创造和发扬光大。

O'Reilly为软件开发人员带来革命性的“动物书”；创建第一个商业网站（GNN）；组织了影响深远的开放源代码峰会，以至于开源软件运动以此命名；创立了*Make*杂志，从而成为DIY革命的主要先锋；公司一如既往地通过多种形式缔结信息与人的纽带。O'Reilly的会议和峰会集聚了众多超级极客和高瞻远瞩的商业领袖，共同描绘出开创新产业的革命性思想。作为技术人士获取信息的选择，O'Reilly现在还将先锋专家的知识传递给普通的计算机用户。无论是通过图书出版、在线服务或者面授课程，每一项O'Reilly的产品都反映了公司不可动摇的理念——信息是激发创新的力量。

业界评论

“O'Reilly Radar博客有口皆碑。”

——*Wired*

“O'Reilly凭借一系列（真希望当初我也想到了）非凡想法建立了数百万美元的业务。”

——*Business 2.0*

“O'Reilly Conference是聚集关键思想领袖的绝对典范。”

——*CRN*

“一本O'Reilly的书就代表一个有用、有前途、需要学习的主题。”

——*Irish Times*

“Tim是位特立独行的商人，他不光放眼于最长远、最广阔的视野，并且切实地按照Yogi Berra的建议去做了：‘如果你在路上遇到岔路口，走小路（岔路）。’回顾过去，Tim似乎每一次都选择了小路，而且有几次都是一闪即逝的机会，尽管大路也不错。”

——*Linux Journal*

译者序

背景

时至今日，Hadoop 已形成了较为成熟、持续发展的生态圈。2016 年是 Hadoop 发展的第十个年头，从 v1 到 v2，再到将要发布的 3.0.0-GA，其功能、性能、稳定性及可用性均得到了极大的提升。Hadoop 在业界和学术界快速地渗透、迭代和普及，已经成为了数据处理领域最为基础的技术选型和基本架构。承担底层分布式存储的 HDFS、经典的分布式计算模型 MapReduce，以及成熟的资源管理任务调度框架 YARN 一同构成了传统概念上的 Hadoop。而基于 Hadoop 的各个组件也在蓬勃发展：进行内存迭代计算的新一代燎原之火 Spark 已至 2.0.2（2016 年 11 月 14 日），能将 SQL 转化成多种执行引擎（MapReduce、Tez、Spark）的 Hive 已至 2.1.0（2016 年 6 月 20），提供键值对存取的多版本海量数据库 HBase 已至 1.2.4（2016 年 11 月 7 日）。Hadoop 生态圈日趋庞大，各种各样的自由软件已逾百种（参见 http://hadoopecosystemtable.github.io）。

Hadoop 凭借开源社区的贡献者和布道师打造分布式环境下的大数据生态。然而，在大批软件不断涌现的背景下，也有一些曾风光无限或崭露头角的项目销声匿迹。开源社区中最为成功的产品非 Linux 莫属。在灵魂人物 Linus Torvalds 的带领下，Linux 进入服务器及嵌入式设备等领域，占领了操作系统的天下，又和 Android 一起占据了移动市场的大半江山，在桌面市场也有所斩获。众多 Linux 发行版大都将常用软件打包在内，并维护有自己的软件仓库，通过版本升级迭代增加新的功能、修复旧的 Bug，这些技术支持和服务都大大提高了 Linux 的易用性。与开源前辈 Linux 类似，Doug Cutting 大牛开创的 Hadoop 运行在廉价商用服务器上，以集群之力，分而治之地解决先前传统数据库、传统存储、传统计算模型束手无策的问题，让大规模数据的处理成为了可能。而 Hadoop 很早就进入了发行版时代，国外的 Cloudera、Hortonworks、DataBricks、MapR、EMC 等公司及国内的华为与星环都推出了自己的定制发行版本，各具特色地打包和修改 Hadoop 生态系统的软件、提供 BugFix 和 Backport 的 Patch，以及对应的增值服务和技术支持，如耳熟能详的 CDH、HDP 等。Amazon 的 EMR、阿里云的 MaxCompute（原 ODPS）、百度的 BMR 也是类似的产品。开源社区与企业有着不尽相同的业务场景和技术路线，单就交互式 SQL 引擎来讲，Cloudera 力推的 Impala 及优化存储 Kudu、Hortonworks 持续优化的 Hive、Spark 生态中原

生的 Spark SQL、Facebook 的 Presto、EMC 的 HAWQ、MapR 的 Drill 以及基于 HBase 的 Kylin 和 Phoenix 都是 SQL-on-Hadoop，但各有千秋。Hadoop 生态呈现碎片化趋势的同时，也有了百家争鸣的氛围。而开源与企业的结合，让 Hadoop 生态良性发展，也让数据唾手可得。作为开发者和用户，想要了解、使用、融入开源，除了各种博客、论坛、会议以及邮件列表之外，阅读文档和代码是不二之选。

关于本书

本书系统地展现了 Hadoop 生态圈的全景图，能够在面向问题解决的各种博客、论坛以及邮件列表之外，让读者可以在了解了整体架构和基本原理之后更好地去应用和实施，是一本面向体系建设和应用实现的教科书。所谓磨刀不误砍柴工，面向问题解决的思路如同充饥果腹的快餐，常常会急匆匆地解决问题，或不知其所以然、或不小心埋下深坑、或错失了更好的方案；而构建在基础知识和理论之上的架构体系和应用经验则是均衡营养、合理膳食的大餐，带来健康完善的思路、可以少走弯路、规避风险。更可贵的是，本书的第二部分专门介绍了大数据领域常见的三类应用场景，相信可以提高读者拆解业务需求、进行技术选型、更好地实现应用的能力。

致谢

感谢本书的编辑朱巍、岳新欣、谢婷婷、张憬，不辞辛苦地修改我扭曲的文字，让它可见天日。

感谢影响着我人生观价值观的祖父母及父母，让我可以安稳地长大成人，愿他们一世安详、永世安康；感谢我的妻子长久以来的支持和付出，愿我们分享阳光、分担风雨，一同携手走下去。

欢迎大家搜索“Hadoop 应用架构”QQ 群或直接输入“345527351”群号，加入本书的读者 QQ 群，交流大数据的那些事儿，解决大数据相关的各种问题。

是为序。

郭文超

2016年11月20日 于北京

序

在过去的十年中，Apache 软件基金会的 Hadoop 项目蓬勃发展、欣欣向荣。

Hadoop 起初是 Nutch 项目的一部分，旨在提供一种前景远大的功能——通过扩展的方式支持到 PB 级数据的处理。2005 年，Hadoop 最多只能在几十台机器上运行，并且很不完善。只有一小部分人拿 Hadoop 做做试验，练练手。然而，一些人看到了它的前景与趋势，一种经济的、可扩展的、通用的数据存储和处理框架，有着广阔的实用价值。

到 2007 年，Hadoop 的高扩展性在 Yahoo! 公司得到了证实。目前 Hadoop 已经可以在数千台机器上可靠地运行了。Yahoo! 是第一个将 Hadoop 用于产品级应用的公司，而后其他互联网公司也相继使用，如 Facebook、LinkedIn 以及 Twitter 等。虽然这个时期的 Hadoop 在 PB 级的数据处理上拥有良好的扩展性，但考虑到无安全控制和只有 Java 语言的批处理接口，真正使用的话成本较高。

再后来，Hadoop 作为一个复杂生态系统的核心，增加了细粒度的安全控制、组件的高可用性（High Available，HA）以及通用的调度器（YARN，Yet Another Resource Negotiator，另一种资源协调者）。

围绕 Hadoop 这一核心，产生了一大批不同类型的工具。拿 HBase 与 Accumulo 来说，它们都能够提供在线的键值对存储，快速响应交互式的应用访问。而其他一些工具，如 Flume、Sqoop 以及 Apache Kafka，能够帮助完成数据在 Hadoop 存储层上的接入与导出。而 Pig、Crunch 以及 Cascading 提供了增强版的 API。SQL 查询能够通过 Apache Hive 与 Cloudera Impala 处理。Apache Spark 算是一个超级明星，能够提供更强大、更优化的批处理 API，同时还支持实时流处理、图数据处理与机器学习。Apache Oozie 与 Azkaban 能够协调和调度以上工具。

是不是对这些感到有些迷惑呢？叩开 Hadoop 生态系统的大门，一大波工具正汹涌地扑面而来。要想高效利用这一新平台，需要理解这些工具之间是如何适配的，以及哪个对你有所帮助。本书作者在基于 Hadoop 构建应用系统方面均拥有多年经验，本书就是这些经验和智慧的结晶。

理论上来说，有各种各样的方式配置和连接这些工具。但是实际上，工具的使用存在着成功的模式。本书描述了最好的实践经验，讲述每个工具的闪光点，以及怎样才能针对特定的任务发挥该工具的最大作用。另外，本书也提供了一些常见的用例。最初的使用者都没多少经验，尝试过各种工具的整合，但是本书描述了已被反复证明有效的模式，可以为读者节省大量的探索时间。

本书作者提供了使用这一强大平台所需要的基础知识。享受这本书吧，让它帮助你创建优秀的 Hadoop 应用！

Doug Cutting
加州院内棚中

前言

毫不夸张地说，在数据管理和数据处理领域，Apache Hadoop 带来了革命性的进展。Hadoop 的技术能力，使得许多行业中的组织能够解决以往技术不可能解决的问题。这些能力包括：

- 大规模数据的高扩展性处理；
- 不管什么格式和结构（或者缺少结构）的数据，都可灵活地处理。

Hadoop 另外一个值得注意的特点在于，它是一个意在相对廉价的商用硬件上运行的开源项目。相对于传统的数据管理解决方案来讲，Hadoop 在可以接受的成本范围内，提供了高扩展性和灵活性。

强大的技术能力，加上较低的经济成本，使得 Hadoop 及其生态系统中的诸多工具快速发展。而且，活跃的 Hadoop 社区也引入了大量支持 Hadoop 数据管理和处理的工具和组件。

尽管发展迅速，Hadoop 仍是一项相对年轻的技术。许多组织仍在尝试了解如何使用 Hadoop 来解决问题，以及如何将 Hadoop 及相关工具应用到真实场景中来形成解决方案。Hadoop 生态系统包含许多工具、应用编程接口（Application Programming Interface，API）及开发选项，这为开发人员提供了更多的选择余地和更大的灵活性，但也使得选择最佳的工具来实现数据处理应用成为了一项挑战。

我们在与大量客户协作的过程中，在与想要了解如何构建可靠、高扩展的 Hadoop 应用的 Hadoop 用户交流的过程中，积累了一些经验，受此启发编写了本书。本书目标并非为现存的工具提供详尽的文档描述，而是在基于 Hadoop 使用这些工具建设可扩展和可维护的应用架构方面，提供指引。

我们假定本书读者对 Hadoop 及相关工具有一定的经验。读者应熟悉 Hadoop 的核心组件，如 Hadoop 分布式文件系统（Hadoop Distributed File System，HDFS）及 MapReduce。关于 Hadoop 及其核心概念，可参见 Tom White 的《Hadoop 权威指南》，这本书的确文如其名。

下面介绍本书中涉及的一系列比较重要的工具和技术，包括扩展阅读的参考资料。

- YARN

  直到不久前，Hadoop 的核心组件通常被认为是 HDFS 和 MapReduce。随着 Hadoop 中一个处理框架的引入，这种情况迅速发生了改变。由于 YARN 的引入，Hadoop 快速转型成为一个支持多种并行处理模型的大数据平台。YARN 为 Hadoop 数据处理提供了通用的资源管理器和调度器，不仅包括 MapReduce，还支持其他的数据处理模型。这使得在单个 Hadoop 集群上可以支持多个处理框架和多样的工作负载，并使得这些不同的模型和负载可以有效地共享资源。关于 YARN，欲了解更多，可参见《Hadoop 权威指南》或 Apache YARN 官方文档（http://hadoop.apache.org/docs/current/hadoop-yarn/hadoop-yarn-site/YARN.html）。

- Java

  Hadoop 和与之相关的许多工具都是使用 Java 语言编写的，并且许多基于 Hadoop 的应用开发也是使用 Java 语言。虽然面向非 Java 开发者的工具和概念不断涌现，但是对于使用 Hadoop 的用户来讲，了解 Java 仍然是弥足轻重的。

- SQL

  虽然 Hadoop 为数据打开了多种处理框架之门，但 SQL 仍然是 Hadoop 数据查询的常用接口。这是因为大量的开发人员和分析人员熟悉 SQL，所以使用 Hadoop 时了解如何写 SQL 仍然是很有意义的。关于 SQL 的介绍，可以参考 Lynn Beighley 编写的 *Head First SQL*。

- Scala

  Scala 是一种在 Java 虚拟机（Java Virtual Machine，JVM）上运行的编程语言，它支持面向对象编程和函数式编程模型。虽然 Scala 是面向通用场景的编程语言，但是它已成为大数据领域越来越流行的语言了。无论是在与 Hadoop 交互的项目实现中，还是处理数据的应用开发中，皆是如此。使用 Scala 作为基础实现语言的项目如 Apache Spark 和 Apache Kafka。因此，基于 Spark 的应用开发也支持使用 Scala。在本书中，许多示例也是使用 Scala 编写的。如果需要了解 Scala，可参见 Cay S. Horstmann 所著的《快学 Scala》；若想深入了解，请参考 Dean Wampler 和 Alex Payne 合著的《Scala 程序设计（第 2 版）》[1]。

- Apache Hive

  提到 SQL，不得不提 Hive，它是一个用于 Hadoop 数据处理和数据建模的非常流行的工具，提供 HDFS 上数据的结构化定义，及数据的类 SQL 查询功能。Hive 项目中包含有一个元数据存储，它不仅以 Hive 的数据结构来存储元数据信息（就是描述数据的数据），还可供其他组件访问，如 Apache Pig（一个更高一层的并行编程抽象）和 MapReduce，其中后者需要借助 HCatalog 组件。另外，其他开源项目——如 Cloudera Impala，一个 Hadoop 之上的低延迟查询引擎——也可以使用 Hive 的元数据存储服务，该服务可以提供对 Hive 中预先定义的对象的访问。关于 Hive，欲了解更多，请参见 Hive 网站（https://hive.apache.org/）、《Hadoop 权威指南》，或 Edward Capriolo 等人所著的《Hive 编程指南》。

注 1：本书已由人民邮电出版社出版。——编者注

- Apache HBase

  HBase 是 Hadoop 生态圈中另外一个频繁使用的组件。它是一个分布式 NoSQL 数据存储，提供 HDFS 上超大规模数据集的随机访问。虽然被称为 Hadoop 数据库，但 HBase 与关系型数据库截然不同，熟悉传统关系型数据库系统的人要想了解 HBase，需接受新的概念。HBase 是许多 Hadoop 架构中的一个核心组件，本书中多有涉及。欲了解更多有关 HBase 的内容，可参考 HBase 网站（https://hbase.apache.org/）、Edward Capriolo 所著的《HBase 权威指南》，或 Nick Dimiduk 和 Amandeep Khurana 合著的《HBase 实战》。

- Apache Flume

  Flume 是一个常用的数据采集工具，将基于事件的数据（如日志）转存至 Hadoop。我们就 Flume 的最佳实践和部署架构进行了整体总结和细节描述。关于 Flume 的更多细节，可参见 Flume 文档（http://flume.apache.org/documentation.html），或《Flume：构建高可用、可扩展的海量日志采集系统》。

- Apache Sqoop

  Sqoop 是 Hadoop 生态圈中另外一个流行的工具，它用来在外部数据存储（如关系型数据库）与 Hadoop 之间进行数据移动。我们会讨论 Sqoop 的最佳实践，以及在 Hadoop 架构中它的最佳位置。关于 Sqoop 的更多细节，可参见 Sqoop 文档（http://sqoop.apache.org/docs/1.4.5/index.html），或 Apache Sqoop Cookbook（O'Reilly）。

- Apache ZooKeeper

  恰如其名的 ZooKeeper 项目旨在提供一个集中化的服务，用来保障 Hadoop 生态圈中各个项目间的协同工作。本书中提及的大量组件，如 HBase，就依赖于 ZooKeeper 提供的服务，所以对 Zookeeper 有基本的了解是有益处的。参考 ZooKeeper 网站（http://zookeeper.apache.org），或 Flavio Junqueira 和 Benjamin Reed 合著的《ZooKeeper：分布式过程协同技术详解》。

由此可见，本书的重点在于 Hadoop 生态圈中的开源工具。值得注意的是，很多传统企业级软件厂家提供了对 Hadoop 的支持，或者处于添加支持的过程中。如果你所在的公司已经使用了这样的企业级工具，那么尝试将这类工具集成到你的 Hadoop 应用开发环境中是大有裨益的。毕竟完成一项任务最好的工具是先前已经熟悉的工具。虽然了解本书中提及的工具，了解它们是如何集成到 Hadoop 中实现应用的，这些都是有意义的，不过在你的环境中选择使用第三方工具也是一个不错的选择。

重申一下，本书的目标不是介绍具体如何使用各种工具，而是讲述什么时候和为什么使用这样那样的工具，同时介绍最佳实践，以及最佳实践适用时的建议和不适用时的调整方法。我们希望本书能够对你构建成功的 Hadoop 解决方案有所帮助。

示例代码

就本书中的示例代码简单声明如下。我们尽量保证本书中的示例是最新的，并确保其正确性。获取最新版本示例代码，请访问本书的 GitHub 地址：https://github.com/hadooparchitecturebook/hadoop-arch-book。

目标读者

本书面向软件开发人员、架构师及项目主管等，满足大家了解 Apache Hadoop 及生态圈中工具的使用方法、建设端到端数据管理方案、集成 Hadoop 到已有数据管理架构等需求。我们的目标并不是深入研究特定的技术，比如 MapReduce，因为已有其他相关的参考资料。我们的目标是：介绍如何高效地集成 Hadoop 生态圈中的组件，以形成一个从原始数据开始直到数据消费掉的完整数据流水线，以及如何将 Hadoop 集成到已有的数据管理系统之中。

我们假定读者对 Hadoop 及相关工具（如 Flume、Sqoop、HBase、Pig 及 Hive 等）有所了解，但我们也提供了合适的参考资料作为补充内容。我们假定读者拥有 Java 编程经验，以及 SQL 和传统数据管理系统（如关系型数据库管理系统）的使用经验。

如果你是一名拥有 Hadoop 背景的技术专家，想要寻求架构或完整方案设计方面的最佳实践或者示例，那么本书再合适不过了。即使你是一名 Hadoop 专家，我们认为本书基于我们使用 Hadoop 的多年经验，包含了许多指引和最佳实践，仍然会让你有所受益。

通过本书，管理人员可基于实际的目标和项目情况，了解何种技术适用，从而为开发者选择合适的培训。

写作目的

多年以来，我们使用 Hadoop 构建大数据解决方案，无论是作为用户还是做客户支持，积累了一些经验。同时，Hadoop 市场也迅速成熟起来，关于深入了解 Hadoop 这一题材的资料也大量涌现。关于 Hadoop 及生态圈的相关工具，有大量的书籍、网站、课程等。尽管有如此多的资料，但在“有效集成这些工具以形成完整的解决方案”这一主题上，相关资源仍显不足。

与用户沟通时，无论这些用户是我们的客户、合作伙伴，还是与会人员，我们发现了一个共同的现象：在“对 Hadoop 有所了解”与“能够使用 Hadoop 解决实际问题”之间存在着巨大的鸿沟。举例来说，市面上有大量不错的资料可以帮助你了解 Apache Flume，但是如何判断这个工具是否适合你的用例呢？而且，一旦确定选择了 Flume 作为解决方案，怎样才能把它高效地集成到架构中呢？为了高效地使用 Flume，需要知道哪些最佳实践，以及需要做出哪些考量？

本书的目的就是缩小“对 Hadoop 有所了解”与“能够使用 Hadoop 形成实际解决方案”之间的差距。书中会介绍利用 Hadoop 实现解决方案时需要考虑的核心内容，并针对几个常见的用案提供完整的、端到端的解决方案示例。

本书结构

本书内容是按照在 Hadoop 上搭建解决方案的流程组织的，首先是 Hadoop 上的数据建模，接下来是将数据导入和导出 Hadoop，以及数据落地到 Hadoop 之后的数据处理，等等。当然，读者可以按照实际需求跳过部分内容。第一部分主要涵盖了使用 Hadoop 创建应用程

序时需要考虑的问题，包括以下几章。

- 第 1 章的内容是 Hadoop 中的数据存储和数据建模，例如文件格式、数据组织及元数据管理。
- 第 2 章的内容是 Hadoop 上的数据导入和导出。这一章将会讨论数据采集和抽取时需要考虑的问题和模式，包括常见工具的使用，如 Flume、Sqoop 和文件传输。
- 第 3 章介绍 Hadoop 上访问和处理数据的工具和模式。我们会在这一章讨论常见的数据处理框架，如 MapReduce、Spark、Hive 和 Impala，以及各自适合的应用场景。
- 第 4 章通过讲述 Hadoop 上一些常见用例的实现方案来继续讨论数据处理框架。我们会使用 Spark 和 SQL 实现具体的例子，来阐释如何解决常见问题，比如数据去重和时间序列数据的处理。
- 第 5 章主要讨论在 Hadoop 上处理海量图数据的工具，如 Giraph 和 GraphX。
- 第 6 章讨论将各种过程与应用协调和调度工具（如 Apache Oozie）整合在一起。
- 第 7 章讨论 Hadoop 上的近实时处理。我们在这里会讨论相对较新的流式数据处理工具，如 Apache Storm 和 Apache Spark Streaming。

在第二部分中，我们将会在 Hadoop 上端到端地实现一些常见的应用程序。这几章的目的在于提供翔实的案例，讲述如何使用第一部分中提到的各个组件，来实现一个基于 Hadoop 的完整解决方案。

- 第 8 章介绍了一个基于 Hadoop 的点击流分析的示例。对于运行大型网站的公司来讲，点击流数据的存储和处理是一个非常常见的用例。它也适用于处理任何机器数据的应用。这一章中，我们会讨论使用 Flume 和 Kafka 这样的工具进行数据采集，讨论如何高效地存储和组织数据，并展示处理数据的示例。
- 第 9 章介绍了一个基于 Hadoop 的欺诈检测应用的示例，这是 Hadoop 一个日益常用的应用场景。这一示例将会涵盖在欺诈检测的解决方案中如何使用 HBase，以及如何使用近实时处理。
- 第 10 章的案例研究探索的是另外一个常见用例：使用 Hadoop 扩展已有的企业级数据仓库（Enterprise Data Warehouse，EDW）环境。这包括将 Hadoop 作为 EDW 的补充，并提供传统数据仓库的基本功能。

排版约定

本书使用了下列排版约定。

- 楷体
  表示新术语。
- 等宽字体（`Constant width`）
  表示程序片段，以及正文中出现的变量、函数名、数据库、数据类型、环境变量、语句和关键字等。
- 加粗等宽字体（**`Constant width bold`**）
  表示应该由用户输入的命令或其他文本。

- 等宽斜体（*Constant width italic*）
  表示应该由用户输入的值或根据上下文确定的值替换的文本。

该图标表示提示、建议或一般注记。

该图标表示警告或警示。

使用代码示例

补充材料（代码示例、练习等）可以从 https://github.com/hadooparchitecturebook/hadoop-arch-book 下载。

本书是要帮你完成工作的。一般来说，如果本书提供了示例代码，你可以把它用在你的程序或文档中。除非你使用了很大一部分代码，否则无需联系我们获得许可。比如，用本书的几个代码片段写一个程序就无需获得许可，销售或分发 O'Reilly 图书的示例光盘则需要获得许可；引用本书中的示例代码回答问题无需获得许可，将书中大量的代码放到你的产品文档中则需要获得许可。

我们很希望但并不强制要求你在引用本书内容时加上引用说明。引用说明一般包括书名、作者、出版社和 ISBN。比如："*Hadoop Application Architectures* by Mark Grover, Ted Malaska, Jonathan Seidman, and Gwen Shapira (O'Reilly). Copyright 2015 Jonathan Seidman, Gwen Shapira, Ted Malaska, and Mark Grover, 978-1-491-90008-6"。

如果你觉得自己对示例代码的用法超出了上述许可的范围，欢迎你通过 permissions@oreilly.com 与我们联系。

Safari® Books Online

Safari Books Online（http://www.safaribooksonline.com）是应运而生的数字图书馆。它同时以图书和视频的形式出版世界顶级技术和商务作家的专业作品。技术专家、软件开发人员、Web 设计师、商务人士和创意专家等，在开展调研、解决问题、学习和认证培训时，都将 Safari Books Online 视作获取资料的首选渠道。

对于组织团体、政府机构和个人，Safari Books Online 提供各种产品组合和灵活的定价策略。用户可通过一个功能完备的数据库检索系统访问 O'Reilly Media、Prentice Hall Professional、Addison-Wesley Professional、Microsoft Press、Sams、Que、Peachpit Press、Focal Press、Cisco Press、John Wiley & Sons、Syngress、Morgan Kaufmann、IBM