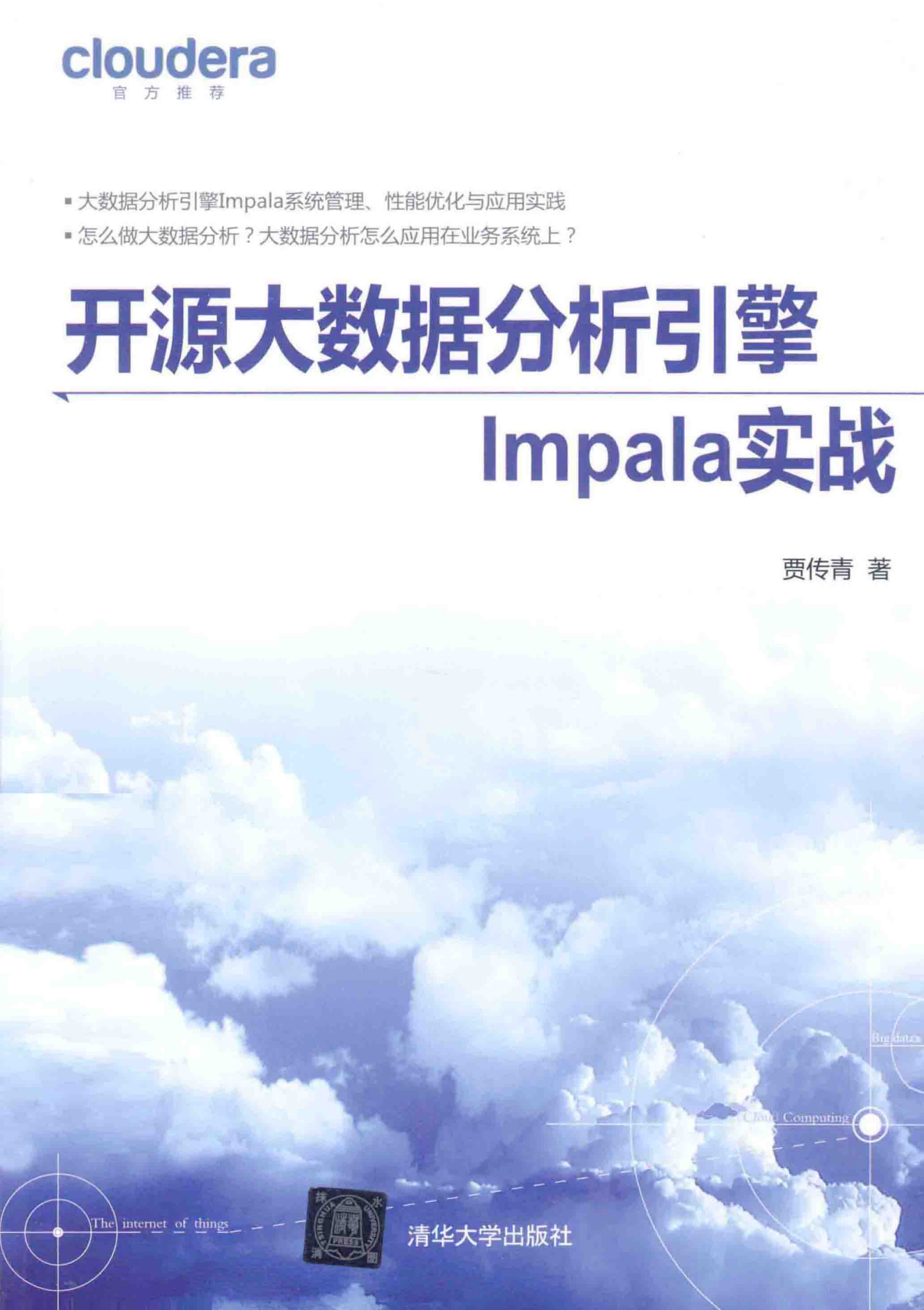
cloudera
官 方 推 荐

- 大数据分析引擎Impala系统管理、性能优化与应用实践
- 怎么做大数据分析？大数据分析怎么应用在业务系统上？

开源大数据分析引擎
Impala实战

贾传青 著



清华大学出版社

开源大数据分析引擎

Impala实战

贾传青 著

清华大学出版社
北京

内 容 简 介

Impala 是 Cloudera 公司主导开发的新型查询系统，它提供 SQL 语义，能查询存储在 Hadoop 的 HDFS 和 HBase 中的 PB 级大数据。Impala 1.0 版比原来基于 MapReduce 的 Hive SQL 查询速度提升 3～90 倍，因此，Impala 有可能完全取代 Hive。作者基于自己在本职工作中应用 Impala 的实践和心得编写了本书。

本书共分 10 章，全面介绍开源大数据分析引擎 Impala 的技术背景、安装与配置、架构、操作方法、性能优化，以及最富技术含量的应用设计原则和应用案例。

本书紧扣目前计算技术发展热点，适合所有大数据分析人员、大数据开发人员和大数据管理人员参考使用。

图书在版编目（CIP）数据

开源大数据分析引擎 Impala 实战 / 贾传青著.-北京 ：清华大学出版社，2015
ISBN 978-7-302-39002-2

I. ①开… II. ①贾… III. ①关系数据库系统 IV. ④TP311.138

中国版本图书馆 CIP 数据核字（2015）第 005181 号

责任编辑：夏非彼
封面设计：王　翔
责任校对：闫秀华
责任印制：刘海龙

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 装 者：北京密云胶印厂
经　　销：全国新华书店
开　　本：190mm×260mm　　印　　张：21.75　　字　　数：557 千字
版　　次：2015 年 3 月第 1 版　　印　　次：2015 年 3 月第1 次印刷
印　　数：1～3000
定　　价：59.00 元

产品编号：057645-01

Cloudera 官方推荐序（中文）

大数据，作为目前工业界的主要技术趋势，定位于转化工业界的每一个细分市场，推动企业运用其数据开展业务的革命，并从根本上改变了支撑现代社会的 IT 基础架构。毫无疑问，大数据对中国意义重大，它给中国 IT 业的创新带来了巨大机会，没有其他任何一个国家比中国有更多的人口、更多的设备和更多的数据。

目前 Hadoop 是用于大数据的优选平台解决方案。作为 Hadoop 技术以及提供 Hadoop 解决方案的领导者，Cloudera 不仅提供经过了业界验证的 Hadoop 平台解决方案，也提供功能强大的工具帮助企业用户充分利用 Cloudera 企业版 Hadoop 解决他们的业务问题。Impala 就是 Cloudera 开发的众多强大工具之一。

Impala 是为了在 Hadoop 上实现低延迟的 SQL 查询而设计开发的，它原生地运行在 Hadoop/HBase 存储系统和元数据之上，因此它继承了 Hadoop 的灵活性、伸缩性和经济性，具有分布式本地化处理的特性以避免网络瓶颈，它与现有 Hadoop/CDH 的、基于工业标准的 SQL 接口兼容。它支持交互式 SQL，比最新版本的 Hive 快很多倍。由于 Impala 的这些优势，它受到了全球企业用户的热烈欢迎。

看到将为中国读者发布的这一本中文版 Impala 书籍，我非常欣喜，这无疑对中国用户更好地使用 Hadoop，解决他们的业务问题有很大帮助。因此，我要感谢所有为发布本书做出贡献的人们。最后，也要感谢广大读者对 Impala 的喜爱，以及你们在大数据——这一令人激动的 IT 发展方向上所做的贡献！

苗凯翔　博士

Cloudera 副总裁

Cloudera 官方推荐序（英文）

Big Data, as the next major trend in the industry today, is set to transform every market segment in the industry, revolutionize how enterprises will do their businesses using their data, and fundamentally shift the IT infrastructure underlying modern society. No doubt, Big Data is of great significance to China and it presents excellent opportunities to the IT industry in China for new innovations - no other country has more people, more devices, and more data than in China.

Hadoop is a platform solution of choice today for Big Data. As a leader of Hadoop and Hadoop-based solutions, Cloudera offers not only solid industry-proven Hadoop platform solutions, but also powerful tools to help enterprise users to fully leverage the Cloudera Enterprise Hadoop to solve their business problems. Impala is such a powerful tool, among many others created by Cloudera.

Impala is purpose-built for low-latency SQL queries on Hadoop, operating natively on Hadoop/HBase storage and metadata. As such it inherited the flexibility, scale, and cost advantages of Hadoop, and features distributed local processing to avoid network bottlenecks, and is compatible with SQL interface for existing Hadoop/CDH application based on industry standard SQL. It supports Interactive SQL which is typically many times faster than the latest Hive. Because of these advantages Impala has, it has become hugely popular among many enterprise users worldwide.

Now, I am very glad to see that a book focusing on Impala will be available in Chinese for the readers in China, which will definitely have a positive impact in helping users in China in better utilizing Hadoop and in solving the business problems they have. For this reason, I would like to thank all who have made contributions in making this book available in local language in China. Finally, I would like to thank the reader for your interest in Impala and for your contributions in this exciting direction of IT which we call Big Data!

Kai X. Miao（苗凯翔），Ph.D

Vice President, Cloudera Corporation

推荐序二

Impala是Hadoop生态圈中不可或缺的一个环节，它提供SQL语义，能够对HDFS和HBase中的PB级大数据进行交互式实时查询，从而弥补了Hive批处理的不足。本书是国内第一本Impala专业书籍，相信对您有益。

刘鹏
中国云计算专家咨询委员会副主任、秘书长
中国信息协会大数据分会副会长

推荐序三

Impala 起源于谷歌的 Dremel 大数据快速分析处理平台论文，由著名的大数据公司 Cloudera 开发并开源，在业内的知名度非常高，它立足于内存计算，从多迭代实时批量处理出发，兼顾数据仓库，是大数据系统领域的基于内存和 SQL 的快速处理分析计算平台。

近几年，大数据在如火如荼地发展着，随着谷歌开源了 Dremel 的论文，基于内存的计算分析技术逐渐走入大众的视野，谷歌为大家展示了一种超越 MapReduce 的数据分析之路，3 秒钟分析一个 PB 的数据已经不再遥不可及。但对于该论文的源代码实现寥寥可数，在这有限的开源代码中的翘楚者当属 Impala 和 Spark，但一直以来，无论是 Impala 还是 Spark，国内相关的文档和讨论都不甚全面。而本书的作者贾传青为大家带来了一本基于他多年的数据库和分布式工作的经验，可以说，这是 Impala 在目前国内最全面、最完整的技术讲解书籍。

随着内存分布式计算技术的发展，目前国内讨论该领域的社区也越来越多，但内存计算技术也存在几个问题，例如，Apache 的 Drill 项目迟迟没有开源。Spark 对于普通用户来说学习成本太高。Impala 则很好地设计了一个相对中庸的方案，在兼顾计算速度的同时，也照顾原有的 SQL 分析师和 Hive 的用户们。对于用户的快速上手开启大数据之旅十分方便快捷且简单实用。可以说，Impala 基于 HDFS、SQL 和 Hive，却又超越了 SQL、MapReduce 和 Hive。而它与经由传统数据库改造而来的分布式数据库如 Greenplum 等又有极大的差别。

这是国内第一本全面讲解 Impala 的书籍，既可以作为想快速搭建基于 Hadoop 的数据仓库的原数据库爱好者们的优秀参考书籍，又可以成为对 Spark 感兴趣的用户的架构理解入门书籍，因为这两种技术虽殊途但同源，Spark 虽效率更好，但架构更为复杂，开发难度也更高；而 Impala 恰恰为用户在复杂和简单之间兼顾了平衡。Impala 在国内的深入介绍和讨论的材料并不多，所以，作为第一本全面介绍 Impala 的书籍，其编写难度也是可想而知的。

非常荣幸能够为传青的这本书籍做序，同时也感谢他让我看到了 Impala 非常有特色的另一面。他是我认识的国内为数不多的研究 Impala 方面的专家，具有多年的数据库管理与开发经验，同时最近几年也在一直研究分布式和实时计算相关的最前沿技术，可以说，这本书既是他历史经验的总结，又是对他自己的一个新突破。

再次感谢传青邀请我为他的新书做序，在大数据的道路上，愿我们和读者们一起加油，向前。

向磊

EasyHadoop 与 phpHiveAdmin 作者

EasyHadoop 社区创始人，eXadoop 公司创始人

推荐序四

大数据应用的日渐增多为社会生活中的许多领域带来了积极变化。基于 Hadoop 的离线计算提供了强大的数据处理能力，但由于 Hive 底层执行使用的是 MapReduce 引擎，仍然是一个批处理过程，因此难以满足查询的交互性要求。能否有一项技术兼顾 DBMS/Hadoop 的混合优势呢？Cloudera 公司主导开发的 Impala 应运而生。有测试表明，Impala 的性能较 Hive 提高了 3~90 倍。

本书是国内第一本系统介绍 Impala 技术细节的书籍，对 Impala 的概念架构、SQL、资源管理、存储、分区进行了详细介绍。作者结合自身多年的 Oracle 和大数据研发经验，对 Impala 性能优化提出了自己的见解：通过数据对比可以看到良好的设计，以使计算性能有巨大提升。文章的最后，还根据设计原则给出了应用的具体案例。

作者贾传青执着于技术并乐于分享，他一直想写一本看着舒服的技术书籍。希望本书能够为有兴趣研究 Impala 的专业人员或学习者有所帮助。

郭刚

慧聪网 CTO

推荐序五

当下各类大数据技术的涌现已经成为了创新的源泉，让更多的想象成为可能。新一代开源大数据分析引擎 Impala 作为 Cloudera 在 Hadoop 领域的重要成员，其开源生态圈正在快速成长，技术应用前景广阔。贾先生与我在多个大数据技术领域有过深入交流，贾先生深厚的技术功底和严谨的钻研精神给我留下了深刻印象。非常高兴能看见贾先生的新著，这是我截至目前看到的第一本阐述 Impala 技术和应用最体系化的中文书籍。不要沉浸于讨论多大规模的数据才是“大数据”，本书将带领读者快速地掌握这个技术，打开大数据时代的窗户。大数据时代已经来临，先知先觉、先行先试者将先受益。

庄伟波
中信证券

推荐序六

鄙人从事电信行业数据库维护多年，和贾先生是多年的工作伙伴，很钦佩贾先生的才华，也很高兴他的新书能够出版。鄙人于 1999 年进入电信行业工作，从事信息化系统的维护工作，主管过本企业全省的核心营业、计费账务、经营分析等系统。拥有多年的 Oracle 数据库维护经验，是国内最早的 OCM 之一。

电信行业通过网格营销、市场监控、经营决策等提升用户体验，保有市场份额。而同时针对海量数据的分析，时效性又成为不容忽视的问题。天下武学，唯快不破，窃以为 IT 系统亦是如此。本书详细讲解了 Hadoop 生态系统中的实时分析引擎 Impala，相信能帮助每位读者快速掌握这一技术。

郭瑜敏
山西联通

推荐序七

我的博士研究方向是在可持续发展和网络组织背景下运用量化评价和数据分析方法建立区域经济发展模型。期间我沉淀了很多当今“大数据”分析过程中经常用到的数据分析和挖掘技能。近几年我负责了邮政业务量收数据仓库管理系统的建设工作。如何发挥数据仓库的存储和运算效率、持续提升技术投入成本的效益一直是我比较关注的问题。

Impala 作为基于 Hadoop 的实时计算技术可以直接通过 BI 产品进行展现，进行数据的查询和展示。在商业领域，如何发挥“大数据”的商业价值，帮助企业形成核心能力还没有形成一个成熟的框架模式。一些运用“大数据”技术的先行者们开展了积极的尝试，传青就是其中的一位专家。他的努力态度，所取得的成果和工作精神值得敬佩。

刁晓纯 博士
中国邮政
《实用数据分析》译者

作者序

写作背景

作为曾经的传统关系型数据库从业者，我们不仅需要了解数据库本身，还需要了解运行数据库的主机，存储数据库数据的仓库，读取数据库数据的中间件以及应用本身的特点。随着硬件的发展以及数据处理的细化，数据库技术从传统的基于磁盘的关系型数据库，向内存数据库、MPP 数据库不同的方向演进，数据库产品也从高大全向单一 RDBMS 吃遍天、短小精悍的方向发展。在架构时，我们必须根据应用的特点选择合适的数据库产品。

自 2009 年开始，笔者开始尝试使用基于 Hadoop 的技术来解决传统数据库无法线性扩展的问题。Hadoop 不能称之为“数据库”，也不能简单地称之为“应用”，而是介于数据库和应用之间的一种既能用于存储和处理数据，又能处理应用业务逻辑的一个混合体，我们通常称之为“数据平台”。Hadoop 虽在本质上解决了磁盘 IO 的扩展问题，但同时由于其基于磁盘（自 Hadoop 2.3 起支持缓存特性），因此对于某些实时性要求更高的任务无能为力，Impala 及其他的基于内存的运算技术应运而生。

Impala 的存储基于 HDFS，运算基于表的统计信息生成执行计划，具备资源管理功能，是最像传统数据库的大数据技术。笔者着手写作本书时 Impala 的最新版本为 1.3.1，而目前已演进至 2.1 版本，在 SQL 语法、安装、扩展性及性能方面进一步增强。

主要内容

工欲善其事，必先利其器，第 1 章手把手地为大家介绍如何离线搭建一个 Impala 环境。有了一个环境之后，我们可以暂时不考虑细节，先尝尝鲜使用一下它。第 2 章介绍如何在 Impala 上进行简单的数据加载、建表、查询等操作。作为 Impala 的管理者，仅仅能够简单使用它是远远不够的。第 3 章系统地介绍 Impala 的架构体系及各组件的作用。第 4 章是为 Impala 的使用者量身定做的，花费比较大的篇幅介绍了 Impala SQL、函数、UDF 等。任何一款数据库都会提供一个命令行工具，方便在没有图形界面的情况下，或者在 Shell 中进行调用，Impala 也不例外，第 5 章介绍 Impala 的命令行工具 Impala-shell。那如何有效地避免硬件资源的过载使用呢？当然是通过资源管理，第 6 章将详细介绍 Impala 的资源管理机制，另外也可以将 Impala 使用 YARN 来进行管理。第 7 章详细介绍了 Impala 底层支持的文件类型，基本囊括了 Hadoop 主流的文件类型。第 8 章介绍了 Impala 的分区机制。第 9 章介绍了 Impala 性能优化

的指导原则，以及优化过程中使用到的各项技术。第 10 章介绍了在企业应用中使用 Impala 时的设计原则及应用案例。

读者对象

- 内存计算技术初学者
- 数据库管理员及数据库开发人员
- Hadoop 及内存计算的运维工程师
- 开源软件爱好者
- 其他对大数据技术感兴趣的人员

致谢

在此感谢 Cloudera 的苗凯翔博士、Deborah Wiltshire、Yale Wang 对本书的认可。感谢我的好兄弟闫猛、付乐庆对我一直以来的鼓励。感谢我曾经服务过的客户们对我的信任。感谢我的家人和朋友们，你们是我不断努力的源动力。

作者简介

贾传青，数据架构师，Oracle OCM，DB2 迁移之星，TechTarget 特约作家，从数据库向大数据转型的先行者。曾服务于中国联通、中国电信、建设银行、PICC 等，目前供职于一家大数据解决方案提供商，致力于使用大数据技术解决传统数据库无法解决的问题。

作者

2015 年 1 月

目 录