HZ BOOKS
华章IT

作者荣获美国政府颁发的“美国杰出人才”称号。大润发中国区董事长、飞牛网首席执行董事黄明端先生与eBay全球零售科学高级总监逄伟先生作序力荐！

将技术与商业需求相结合，深入剖析大数据商业应用中的困惑与难题，帮助读者更好地掌握技术支撑业务高速发展的方案！

大数据技术丛书

Using Big Data to Build Your Business

大数据架构商业之路

从业务需求到技术方案

黄申◎著

机械工业出版社
China Machine Press

大数据技术丛书

Using Big Data to Build Your Business

大数据架构商业之路

从业务需求到技术方案

黄申◎著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

大数据架构商业之路：从业务需求到技术方案 / 黄申著 . —北京：机械工业出版社，2016.4
（2016.7 重印）
（大数据技术丛书）

ISBN 978-7-111-53528-7

Ⅰ. 大… Ⅱ. 黄… Ⅲ. 商业管理 – 数据管理 Ⅳ. F712

中国版本图书馆 CIP 数据核字（2016）第 078609 号

大数据架构商业之路：从业务需求到技术方案

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：佘 洁 责任校对：董纪丽
印 刷：北京文昌阁彩色印刷有限责任公司 版 次：2016 年 7 月第 1 版第 2 次印刷
开 本：186mm × 240mm 1/16 印 张：19.75
书 号：ISBN 978-7-111-53528-7 定 价：69.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

Foreword 推荐序一

大润发是1998年成立的，到了2008年已是中国最具规模的大卖场，那时候电子商务刚在萌芽阶段，而实体店也还在快速发展，加上2010年公司忙于筹备上市，准备于2011年在香港挂牌，所以我们并没有花太多时间研究电子商务，而且在那个时间段大部分电子商务公司都处于亏损状态。

后来我们惊觉电子商务已快速发展，办公室很多同事都开始在网上购物了，加上“双11”的天量交易额，逼得我们不得不认真地研究电子商务的发展趋势。到2012年，我们发现电子商务越做越好，尤其进入移动互联网时代后，想要满足顾客随时随地的购物需求，电商发展必然是未来消费的新渠道与趋势。所以我们在2013年决定进军电子商务并成立飞牛网（Feiniu.com）。

经过两年多的实践，我认为实体零售跟电子商务还是存在很大差异，其中最主要的差异有两点。

- 顾客忠诚度差异：对于线下卖场而言，选址是关键。地点位置正确，就会有稳定的客流，也容易培养顾客忠诚度。然而，对于线上而言，它不受地域的限制，顾客切换不同的网站是一件十分轻松的事情，因此忠诚度普遍不高。
- 顾客行为数据的获取成本差异：线下卖场很难跟踪顾客的行为，如果要安装各种复杂的信息采集设备，运营成本就会很昂贵。而到了线上，顾客浏览网站时“凡走过必留下痕迹”，电商要收集顾客的行为只需要读取站点的访问日志，可以说相对容易。

正是因为顾客忠诚度不高，对于忠诚顾客的培养成为电商的必争之地。我相信，要实现这个目标，基于顾客行为的大数据和精准化营销就显得更为重要。我们需要充分利用数据挖掘，并快速反馈到整个电商系统。

至于如何做到个性化的搜索和推荐，如何做好客户关系管理（CRM），以及如何做到精准的推送和营销，一直是我们探索的内容。飞牛网从成立之初到现在，碰到了很多与搜索和大数据相关的问题和困难，一年半以前，黄申博士加入了飞牛网的技术团队，他的技术和经验对于我们的帮助很大，在他的指导下我们快速建立了专业的搜索、推荐及用户画像系统。这些都是我们分析顾客、理解顾客、提升顾客在线体验的核心，使得飞牛网和行业先锋之间

的距离在短时间内大幅缩小。关注飞牛的读者，你们可以到飞牛网体验一下个人喜欢的商品，然后你就能细细品味到我们搜索、推荐等大数据相关的功能给你带来的便捷和惊喜。

当然，这些成绩和黄申博士丰富的业界经验分不开。在日常的工作中，他总是有独到的见解。如果你有幸阅读本书，一定能从他的分享中了解大数据是如何运作的，了解大数据是如何支持业务的，以及了解技术是如何满足业务需求的。对于还处在大数据摸索中的人而言，他的思路和探讨非常宝贵，这是一本讲述搜索和大数据领域实践经验的好书，值得推荐。

飞牛网 CEO 黄明端

2016 年 3 月

Forword 推荐序二

最近的十年中，我一直在 eBay 从事数据相关的项目，领导了包括零售科学和搜索科学在内的研发团队。如今 eBay 在全球已有 1 亿多注册用户，每天都有数以百万的家具、收藏品、电脑、车辆在 eBay 上被刊登、贩售、卖出，每年的营业额高达数千亿美元。

但是我们非常清楚，对于 eBay 而言，更为珍贵的财富是网站上每时每刻都会产生的海量数据。通过对这些数据的分析，我们可以指导卖家进行更好的搜索引擎优化、制定更好的价格、控制合理的库存；还可以帮助买家找到更合适自己的商品和更优质的服务。当然，大数据的分析还能帮助 eBay 有效地防范作弊和欺诈，保证整个平台和渠道的健康发展。

正是意识到数据的关键性，eBay 非常重视挖掘和利用它们的潜在价值。本书作者曾经在 eBay 的研究院和搜索科学部门工作，专门从事机器学习的研究和应用。他协助 eBay 构建了数项核心算法及其相关产品，包括基于机器学习的搜索排序、高质量用户评价的发现和摘要、相似和相关商品推荐栏位等。在此过程中，他和各个技术同仁、产品经理、业务部门紧密合作，而这本著作就融入了作者在这些实战项目中所积累的丰富经验。所以，本书最大的闪光点在于，它的内容不仅仅局限于技术本身，而是考虑到了在不同的应用场景下，这些技术应该怎样合理运用。例如，对于基于学习的搜索排序，通常要考虑哪些因素以及怎样的学习模型？对于智能推荐的栏位而言，相似和相关商品又有怎样的区别？分别都应该使用怎样的推荐模型？

除了与业务应用紧密结合，此书还具有覆盖面广和通俗易懂的特点。全书涉及的主题包括大数据的获取、存取、处理、检索、挖掘和评估中的多数主流技术。同时，作者从自己独特的视角出发，对深奥的技术进行了深入浅出的阐述，大幅降低了大数据知识理解的难度。因此，本书也非常适合大数据产品设计者、产品经理或者架构师进行阅读。我相信，对于希望利用大数据解决业务痛点的读者而言，此书是不可或缺的良师益友。

eBay 全球高级总监　逢伟

2016 年 3 月

前　言 Preface

为什么要写这本书

李克强总理提出“大众创业，万众创新”。在如此美好的大环境下，互联网创业如火如荼。各种模式的O2O，各种精彩的移动App，突然之间都冒了出来，正所谓“忽如一夜春风来，千树万树梨花开”。而在其中，大数据因为蕴含着巨大的商业价值，成为这个时代的趋势之一。众人都希望利用好这个“魔棒”，为自己的事业开疆扩土。可是，就笔者在业界的经历来看，真正能挖掘大数据潜力的公司少之又少。笔者一直很好奇，中国的相关人才如此之多，商业市场又如此之大，何以至如此境地呢？为了找到答案，笔者阅读了不少观察性文章，也走访了一些业内的从业者，发现目前的一大窘境是：大数据技术、产品和商业的结合度还远远不够。导致这个现状的原因有很多，具体分析主要有以下几点：

- 涉及范围广：“大数据”本身是一个比较抽象的概念，任何关乎大规模数据的处理，都可以称为“大数据”。因此它既包括了很多已有的技术，如数据挖掘、机器学习、商业智能等，又包括了近几年诞生的新技术[⊖]，如NoSQL相关的生态系统。而且，一个商业需求也可能会涉及多个相关技术。
- 技术含量高：数据挖掘和机器学习之类的算法和大规模数据处理的架构，相对于普通的应用开发而言，需要更多的理论知识和实践经验积累。而商业价值的挖掘程度却往往取决于使用的技术深度。越是钻研得深入，所产生的价值就会越大。
- 发展速度快：最近几年，算法方面有不少的创新，如深度学习（Deep Learning）；系统架构也在不断升级，如Hadoop的第二代框架Yarn、Storm、Spark等实时流式计算，技术的更新换代非常频繁。但是，商业的发展需要技术系统能够随时应变，快速响应，这与技术的飞速发展本身又存在冲突。
- 成熟方案少：大数据的技术多数是免费的，这对于盈利模式而言无疑是有利的，不过代价就是存在一定的稳定性和易用性问题。现在有一些大型的技术公司提供了更

⊖ 请注意，这里的“新”是相对的，互联网和IT技术发展之快，令人咋舌。

成熟的解决方案，但是价格不菲，对于经费并不宽裕的初创公司而言选择余地太少。

以上这些因素都会形成进入大数据领域的门槛，而高门槛势必会导致大数据在工业界应用的步伐放缓。为了解决这个问题，企业需要培养自己的复合型人才，要求业务人员懂技术、技术人员懂业务。只有如此才能让公司使用合适的工具、获得准确的数据、制定合理的方案。

然而，激烈的市场竞争，膨胀的用户需求，不会给创业公司太多的时间去挥霍。在黑夜之中不断摸索的人们，需要明灯指引前进的方向。虽然目前市面上已有一些相关图书做了不错的尝试，但是它们大多数偏向两个极端：一端是面向金融、经济、社会和管理类等非技术型读者，讲述概念、定义、背景和业界的成功案例等；另一端是面向程序员、算法工程师、架构师和数据科学家等纯技术型读者，讲述具体的技术框架、编程范例、系统调试等。能同时覆盖两者的图书可谓凤毛麟角。因此，笔者萌生了通过一本书来帮助企业快速地建立复合型团队，将合理的业务需求尽快转化为实际产品的想法。笔者在写作过程中，力求：

- 易读易懂。通过生动的案例和形象的比喻来解读难点，降低技术理解的门槛。这样就能够让偏向业务的人员更容易理解大数据背后的运作原理，促进他们和技术人员的沟通及协作。
- 可实践性强。通过分享需要大量实践才能积累的宝贵经验，最大程度地针对业务需求和技术方案之间的空白进行弥补。这将有利于技术人员针对不同的业务需求，规划更为合理的技术方案。

本书通过讲述一个虚拟的（如有雷同纯属巧合）互联网 O2O 创业故事，逐步展开介绍各个阶段可能遇到的大数据课题、业务需求，以及相对应的技术方案，甚至是实践解析。让读者身临其境，一起来探寻大数据的奥秘。对于想进一步深入研究技术实现细节的读者，也给出了继续阅读的方向和指导性建议。笔者衷心希望，无论是技术专家、产品经理，还是业务人员，只要阅读了本书便都能愉快地遨游在大数据的海洋中。

读者对象

根据本书撰写的起心动念，笔者觉得其内容适合如下读者：

- 中小互联网创业公司的 CIO、CTO 和技术骨干。他们可以获知常见的互联网公司从创业初期到中期这个阶段里，数据平台需要满足怎样的业务需求（当然，也包括业务方和产品经理所说的“XXOO”了），技术上通常会面临哪些挑战，以及如何解决。
- 中小互联网创业公司的产品经理和项目经理。个人认为，在不久的将来，最炙手可热的产品经理或项目经理一定是懂一些技术的。技术背景将帮助产品经理和项目经理更好地理解哪些是技术上可以实现的，如果可以实现又大致需要多少开发资源。此外，本书所提及的案例也许能提供一些产品设计上的灵感和启发。
- 中小互联网创业公司的 CEO、合伙人。读懂这本书，CIO、CTO 和产品 VP 的招募，

不用靠第三方和人力资源，因为你可以自己来选。这绝对可以帮助公司少走弯路，加速发展。

- 刚刚起步的算法和架构工程师。很多刚刚毕业或工作没多久的朋友，学了一身本领，对新技术也很有热情，苦于没有太多实践的机会。书中的故事浓缩了不少业界实践的经验和心得，如能融会贯通对他们将很有裨益。同时，覆盖面较广的技术课题概述也为他们继续深入研究提供了方向和指导。
- 梦想家。最后的最后，本书也献给那些希望通过大数据技术进行互联网创业的人们。也许现在你既不是“CXO”（CEO、CIO、CTO、CPO、COO 等的统称），也不是产品经理或项目经理，可是你有自己的创业梦想，那么这本书也献给你。

当然，由于侧重点不同，因此本书并不适合钻研技术细节的程序员和编程专家，不过仍然可以在书中找到重要的参考图书指导。同时，本书也不适合关注宏观行业发展的商务人士。

如何阅读本书

为了达到深入浅出、通俗易懂的效果，本书的第一大部分概述了大数据的主要技术，包括大数据的获取、存储、处理，还有架构设计的基本理念，以及常用的消息和缓存机制。这一部分你会发现关于 Nutch、Flume、Hadoop、HBase、Redis、Hive、Kafka、Spark、Storm 等的简介。对于数据处理的高级技术，本书着墨不少，但不乏对于信息检索和数据挖掘课题的探讨。例如站内搜索引擎、推荐系统、广告系统、聚类、分类和线性回归等。由于商业需求尤其看重实际产出，因此第一部分的最后还会分析常见的效果和性能评估。相信这部分对于构建读者的大数据知识体系会很有帮助。在每一章的最后，我们还会给出重要的参考图书，以便于读者继续深入学习。

第二大部分的每个章节都是从业务需求的描述入手，然后进行需求分析，根据需求的特点，对第一大部分所涉及的备选技术进行筛选，最后是技术方案和架构的确定。不同的商业需求可能会使用类似的技术点。但是具体使用方式不会雷同，根据不同的数据集合、不同的应用场景和不同的进阶难度，我们为读者提供了反复温习和加深印象的机会。

勘误和支持

正如前文所述，大数据发展得实在是太快了。可能就在你阅读这段文字的同时，又有一项新的技术诞生了，N 项技术升级了，M 项技术被淘汰了。再加之笔者的水平有限，编写的时间也较仓促，书中难免会出现一些不够准确或有遗漏的地方，不妥之处在所难免，恳请读者通过如下渠道积极建议和斧正，我们很期待能够听到你们的真挚反馈。

QQ：36638279

微信：18616692855
邮箱：s_huang790228@hotmail.com
LinkedIn：https://cn.linkedin.com/in/shuang790228

扫一扫就能联系作者：

公众号：

致谢

首先要感谢上海交通大学尤其是俞勇教授，你们给予我不断学习的机会，带领我进入了大数据的世界。同时，感谢阿里云的高级总监薛贵荣，你的指导让我树立了良好的科研态度。

还要感谢微软亚洲研究院、eBay 中国研发中心、沃尔玛 1 号店、大润发飞牛网和 IBM 中国研发中心，在这些公司十多年的实战经验让我收获颇丰，也为本书的铸就打下了坚实的基础。

感谢曾经的微软战友陈正、孙建涛、Ling Bao、曾华军、张本宇、沈抖、刘宁、严峻、曹云波、王琼华、康亚滨、胡健、季蕾等，eBay 的战友逄伟、王强、王骁、沈丹、Yongzheng Zhang、Catherine Baudin、Alvaro Bolivar、Xiaodi Zhang、吴晓元、周洋、胡文彦、宋荣、刘文、Lily Yu 等，沃尔玛 1 号店的战友韩军、王欣磊、胡茂华、付艳超、张旭强、黄哲铿、沙燕霖、郭占星、聂巍、邵汉成、张珺、胡毅、邱仔松、孙灵飞、凌昱、王善良、廖川、杨平、余迁、周航、吴敏、李峰等，大润发飞牛网的战友王俊杰、陈俞安、蔡伯璟、陈慧文、夏吉吉、文燕军、杨立生、张飞、代伟、陈静、赵瑜、李航等，IBM 的战友李伟、谢欣、周健、马坚、刘钧、唐显莉等。要感谢的同仁太多，如有遗漏敬请谅解，很怀念和你们并肩作战的日子，你们让我学到了很多。

感谢机械工业出版社华章公司的编辑杨绣国（Lisa）老师，感谢你的魄力和远见，在最近的 3 个月中始终支持我的写作，你的鼓励和帮助引导我顺利地完成了全部书稿。也要感谢凌云为我引荐了如此优秀的出版社和编辑。

衷心感谢大润发、飞牛网董事长黄明端先生和 eBay 全球高级总监逄伟先生，在百忙之中为本书作序。也衷心感谢欧电云的董事长韩军先生、永辉集团电商总经理黄志雄先生、美的集团电商总经理吴海泉先生、百度 LBS 新业务产品总监王欣磊先生、阿里巴巴高级产品专家张旭强先生、LinkedIn（领英）的商务分析经理 Yongzheng Zhang 先生、京东商城推荐

搜索部总监刘尚堃先生和唯品会云计算高级总监诸超先生为本书撰写推荐语。

还要感谢我的爸爸、妈妈、岳父、岳母，感谢你们对我写书的理解和支持。

最后我一定要谢谢我的太太 Stephanie 和宝贝儿子 Polaris，为了此书我周末陪伴你们的时间更少了。你们不仅没有怨言，而且时时刻刻为我灌输着信心和力量，感谢你们！

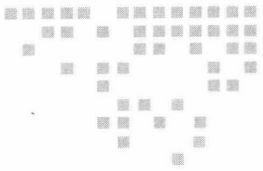
谨以此书，献给我最亲爱的家人，以及众多热爱大数据的朋友。

黄　申

美国，硅谷

2016 年 3 月

Contents 目　录



第 1 章 Chapter 1
抉　择

上海，又是一个春天，阳光透过薄薄的窗帘，懒懒散散地洒入屋内。当一缕光线偷偷地爬上杨大宝的眼角时，他睁开了朦胧的双眼。

等等！杨大宝是何许人也？

杨大宝，姓杨名大宝，土生土长的上海人，从小就喜欢玩电子产品，大学专业是计算机科学，酷爱信息技术和互联网。自从大学毕业后，他就一直任职于一家大的 IT 公司。最近，他面临人生的一项重大选择。原来，有几位志同道合的朋友想拉他一起开创自己的公司。大宝很清楚，这几年中国迎来了创业的黄金时代。李克强总理提出的“大众创业，万众创新”，明确了政策对创业的大力支持。而老百姓的生活水平也在不断提高，各方面的需求也在不断增加，同时各种风险投资也非常充裕。在这样的大背景下，大家的创业热情空前高涨，尤其是互联网，简直可以用“疯狂”来形容。大宝觉得这正是一个实现自己梦想的好契机。不过，放弃目前优厚的薪资待遇和受人尊敬的公司职位，和几个小伙伴去闯荡江湖，也是要冒不少风险的，最终是否能成功也充满了变数，这样做到底值得吗？大宝这几天夜不能寐，晚上做梦也要纠结一番。若不是淘气的阳光溜进来，可能他还要继续在梦里思考。

洗漱完毕，大宝一边吃着早餐，一边接着梳理思路。首先，创业的点子是不错的，主要思想是做线上线下 O2O（Offline to Online）的社区商业模式：将大型社区周边的各种服务行业进行线上化，让用户足不出户，就可以叫外卖、订座、享受美甲、按摩等服务，还可以购买商品。用户的生活需求得到更大程度的满足，商家也可以吸引到更多的线上客流，而公司的平台也能从双方的交易中获得收益，形成多方互赢的局势，市场前景一片光明。其次，因为大宝是团队里唯一懂 IT 技术的骨干，那么公司里整个庞大的网络系统架构肯定会由他来负责。这几年的工作经历让他也积攒了不少设计和开发的实战经验。后端如数据库、ERP（Enterprise Resource Planning）系统、图片服务器，前端如会员注册、购物流程、页面展示

等大宝都有很深入的了解。不过他还是隐约觉得缺了些什么。

吃完了早餐，大宝熟练地打开电脑，开始飞快地在网上查阅资料，钻研成功的互联网站点是如何设计和架构的。就这样，时钟滴滴答答，不知不觉一天过去了。随着夜幕的降临，望着窗外柔和的街灯，大宝深深地吐了一口气，“还缺一个关键词：大数据”，这是他一天研究下来的结论。

等等？大数据又是什么？

好问题，其实此刻大宝心里也没谱，但是他看到好多资料都反复提到这个词。他隐约觉得，如果没有摸清这点，对于这个初创公司而言，就会存在很大的不确定性。可是，目前创业的团队也很多，竞争相当激烈，从来都不缺好的创意，就看谁首先能做得出、做得好、做得快。没有太多的时间留给大宝了。那该如何是好呢？突然，大宝想到一个人，也许能为他解决心中的这个疑惑。

此人就是黄小明，是大宝的表哥。他是知青子女，从小随父母到武汉生活和读书，到16岁的时候回到上海，考入了知名的高校，并且获得了计算机科学的博士学位，可谓知识渊博。毕业后他在几家世界知名的互联网和电子商务公司任职，有十多年的科研和开发经验，目前正在带领团队攻关几个核心项目。

终于，在一个美好的周末下午茶时间，大宝约到了小明。大宝开门见山，针对自己目前的状况和思考的问题进行了说明。

“嗯……大宝，大数据的确是一个非常重要的领域，而且想要上手也有一定的难度。”

“哦，为什么呢？”

“大数据入门的门槛比较高，原因有几点：知识面非常广，技术含量也比较高，此外发展和更新的速度也快得惊人。更为关键的是，这些技术一般都是开源的，很多都需要自己摸索和积累。除非你们考虑直接使用一些大公司比较成熟的付费方案。”

“嗯，如果是创业起步阶段，我们肯定是不会考虑昂贵的商业解决方案的。”

“那问题就更复杂了……不过……”

“不过什么？”

“如果你肯花些功夫来学习，或许我能给你一些建议和启发。”

“哈哈，小明哥，搞了半天你是要自卖自夸啊！”

“这都被你看出来了。其实我最近正在整理这些年的心得体会，准备出版一本关于大数据的书，以便于团队的培养及业界的交流。那我借此机会，先和你讲讲，如何？”

“哇，那求之不得啊！”

“大数据其实是非常宽泛的概念，这里我强调的是如何获取海量的数据，并对它们进行有效的存储、处理和分析，最终让其服务于我们的业务需求。首先，要知道数据的来源非常关键，没有数据就没有生产的原材料。所以我考虑先阐述什么是站外和站内的数据收集系统，以及哪些开源工具可以帮助我们。对于收集到的数据，在第一时间我们要存储它们，然后介绍最近流行的分布式存储系统，确保辛辛苦苦采集而来的数据不会丢失。并说明对于数

据，可以进行哪些基本的处理，以便产生我们所期望的一些数字统计、内容转换等结果。当然，还有很多高级的技术能够让数据产生更大的价值，如信息检索和数据挖掘。接着就是信息检索领域了，包括搜索、推荐、广告等应用。对了，数据挖掘的概念、基本流程和机器学习的主要算法也很重要。有了这些基础之后，还要考虑算法、模型等处理的效果和性能问题，衡量其是否能达到设计的预期。最后将上述知识点串起来，给出全局的概览框架。你看，这样的逻辑顺序你能理解吗？”

“其实，都不太懂……你还是现在就开始教我吧，从你刚说的最基础的开始吧！”