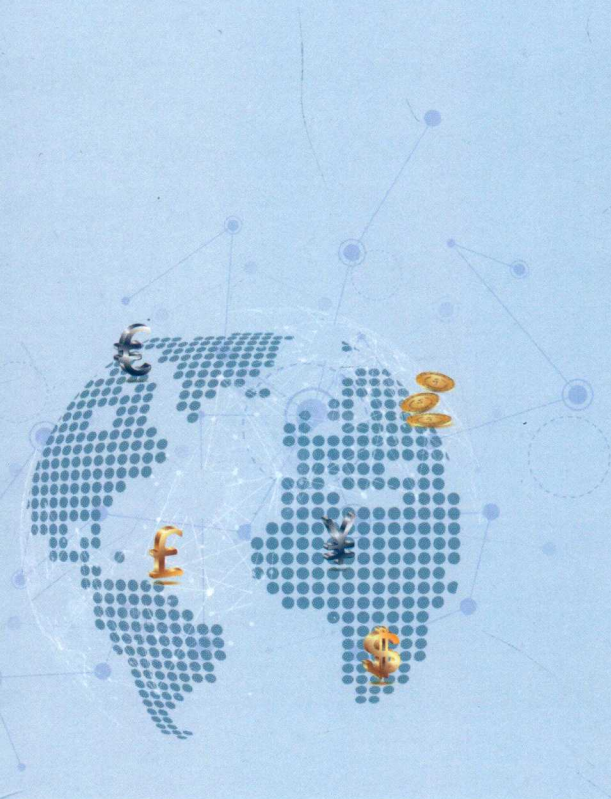


大数据

浪潮之巅

新技术商业制胜之道

徐飞 / 著



中国工信出版集团



电子工业出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
<http://www.phei.com.cn>

大数据

浪潮之巔

新技术商业制胜之道

徐飞 / 著



电子工业出版社

Publishing House of Electronics Industry

北京 • BEIJING

内 容 简 介

本书以各个企业在大数据浪潮中跌宕起伏的经历为核心来讲述大数据发展史，并分析各个大数据企业迥异的发展历程，探讨在新技术浪潮来临时应该如何应对。

本书主要分为两部分，前半部分讲述谷歌、微软、IBM、雅虎、亚马逊、阿里巴巴等大公司在大数据浪潮中的发展史，后半部分讲述各个大数据创业公司的发展历程和现状。在每部分的最后，还通过专门的文章分析并总结了各企业在大数据浪潮中的作为和选择所产生的影响。全书从公司的视角出发为大家呈现了一幅波澜壮阔的大数据领域发展史，读者不仅可以了解大数据技术，更能领略大数据领域的全貌，从各公司的故事中吸取教训，学习思路。

本书适合对大数据技术和商业思维有兴趣的读者阅读。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有，侵权必究。

图书在版编目(CIP)数据

大数据浪潮之巅：新技术商业制胜之道 / 徐飞著. —北京：电子工业出版社，2019.3

ISBN 978-7-121-35834-0

I. ①大… II. ①徐… III. ①数据处理—研究 IV. ①TP274

中国版本图书馆 CIP 数据核字 (2018) 第 296078 号

策划编辑：张春雨 王中英

责任编辑：张春雨

印 刷：三河市双峰印刷装订有限公司

装 订：三河市双峰印刷装订有限公司

出版发行：电子工业出版社

北京市海淀区万寿路 173 信箱 邮编：100036

开 本：880×1230 1/32 印张：7.625 字数：170.8 千字

版 次：2019 年 3 月第 1 版

印 次：2019 年 3 月第 1 次印刷

定 价：69.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：(010) 88254888，88258888。

质量投诉请发邮件至 zltts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

| 推荐语

近年来大数据技术迅速发展，随着智能手机、可穿戴设备以及物联网设备的普及，我们正步入一个数据爆炸的时代，收集、存储和处理大数据的云计算平台应运而生。有了云计算作为基础，人工智能技术才得以迅猛发展，从而更精准地挖掘和释放数据中潜在的价值。

与飞总相识多年，每次与他交流都让我受益匪浅，总能被他在大数据领域的广阔见识和独到见解吸引。飞总的微信公众号“飞总聊 IT”在业界小有名气，他在公众号中连载的“大数据那些事”系列文章是本书内容的基础。本书以浅显易懂的语言分析了各公司的大数据之路，介绍了各个大数据技术平台的功能与特点，内容连贯完整，既能让读者系统了解大数据领域的发展历史和现状，又能促使读者一起思考、交流大数据领域未来的发展方向。

刘国伟

谷歌资深软件工程师

作为一名大数据领域的资深专家，飞总在本书中用生动活泼的语言深入浅出地描绘了大数据生态圈的“众生相”。是什么吸引着我在凌晨时分困意全无地一口气读完了全书？我想，一方面是飞总清晰地描绘了各个公司及其技术相生相克的关系，让我这个实用主义者第一次从宏观角度看到了这个领域的全貌；另一方面，也许对我来说是更重要的方面，书中的内容激发了我对决策和管理的思考：初创公司应该如何平衡短期增长和长期留存？先发优势应该如何保持？后起之秀应该怎样弯道超车？顶尖技术对于商业成功和项目成功的贡献各有多大？大公司将代码开源的行为是在汇集资源还是在让资源流失？管理僵化和混乱哪个会导致公司更快地垮掉？公司应该如何利用资源帮助一个好点子成长，而不是用已有的技术框架绑架它？

所以，即使你对大数据技术并不了解，这本书中丰富的商业案例也会为你带来很多启发。

殷培

Facebook 经理

数据库从业人员往往都保持着大量的输入，读学术论文、读技术总结、读商业分析，每天都要吸纳更多的知识和技能。可是，像飞总这般乐于输出、甘于输出、不计代价地输出的人，倒不常见。这是一种能力，高屋建瓴地梳理归纳，而后有条理、有章程地输出的能力；更是一种态度，已成名的大家总是过于爱惜自己的羽毛，不愿轻易公开地表达对事物的想法和见解，特别是当自己身上已打上公司 tag、

立场 tag，而不是所谓“独立大 V”时。

飞总在本书中并没过多地着墨于技术本身，而是站在技术之上，从宏观的、大环境的层面对商业案例进行分析探讨，在我看来，颇有黄仁宇老先生“大历史观”的感觉。大数据的时代已经开启，大数据的时代远未结束，希望能看见飞总的更多好文，也谢谢他高质量的、持续的分享。

龚玮薇

Oracle 数据库系统技术经理

飞总的文章富有批判精神且论述有理有据，这一点深深吸引了我。一篇篇文章恰似坊间故事，把来龙去脉讲得非常有趣，更于有趣之中蕴含着思考，于思考之间洋溢着情感。愿飞总持之以恒，不断为粉丝们呈现佳作。

李海翔（那海蓝蓝）

数据库内核开发者，

《数据库查询优化器的艺术：原理解析与 SQL 性能优化》、

《数据库事务处理的艺术：事务管理与并发控制》作者，

腾讯 T4 级工程师

我们处在技术高速发展的时代，而商业是技术最好的驱动力。一段代码可能只能默默无闻地放在我们自己的测试服务器上，也可能每天被调用上百万次。决定代码生命周期的往往不是代码质量本身。新技术可能推动商业模式的发展，也可能让项目走向低迷。飞总对技术和商业的关系有着深入的思考，他把这些思考的精华写在了这本书中，将我带入了另一个认知维度。

林晓斌

极客时间“MySQL 实战 45 讲”专栏作者，MySQL 专家

硅谷之所以被称为 IT 圣地，不仅因为这里实现了大量科技创新，更因为这里实现了技术和商业的完美结合，值得国内的 IT 企业认真思考和学习。徐飞的这本《大数据浪潮之巅：新技术商业制胜之道》以通俗易懂、幽默风趣的语言，将硅谷和国内典型的技术与商业融合创新的案例娓娓道来，其中一些是失败的教训，比如 MongoDB 追求技术上的易用性而忽视了商业应用的安全性，因而给了微软的 CosmosDB 大举侵占市场的机会；另一些则是成功的经验，比如 Spark 通过综合不同大数据存储系统的优点，并针对性地改造其缺点实现了后来居上。通过介绍这些正反面案例，带领读者从中汲取经验，学习思路。

谢梁

滴滴首席数据科学家

大数据是被称为“未来石油”的珍贵资源，大数据的概念一旦被提出，便成为各大知名 IT 企业争夺的制高点。谷歌、微软、IBM、阿里巴巴等公司都不惜重金进行相关技术的研究和商业化实践，一时间大数据江湖风起云涌。

飞总在微软、Tableau、华为西雅图研发中心等多家公司从事过十余年大数据相关的工作，对大数据技术以及商业化实践形成了独到而深刻的见解。本书用浅显易懂又风趣幽默的语言，向读者分析了 MongoDB、微软、谷歌、亚马逊、Facebook 等世界 IT 巨头及各个创业公司的技术革新、商业模式变化，以及由此导致的企业沉浮。既能满足读者的猎奇心理，又能让人学到大数据、技术型企业的商业运作等专业知识，给人带来了深刻的思考与丰富的启迪。

汪云海

山东大学计算机学院教授

| 推荐序一

第一次读到徐飞的大数据系列文章是在他的“飞总聊 IT”公众号上，我一口气读完了所有历史文章。后来有机会见到徐飞，才发现我们以前是同事。我于 2006 年加入微软的必应搜索引擎开发部门，主要从事搜索技术研发；徐飞比我晚三年加入，他当时直接进入了必应的大数据 Cosmos 部门。说来也奇怪，我们在同一栋楼工作了若干年，却没有正式见过面，这可能是因为那段时间必应正野蛮生长，人员流动比较大。

过去十年，整个互联网产业发生了翻天覆地的变化，从大数据到云计算再到人工智能，变革一浪接着一浪。读徐飞的大数据系列文章，我再次感受了这场轰轰烈烈的变革。读完文章再看现在的格局，有种以史为镜的感觉，对现在的很多挑战都有了新的思考。作为一名大数据老兵，徐飞是科班出身，所写的文章不仅有精辟的技术见解，而且穿插了很多背后的故事，引经据典、娓娓道来，有种读小说的感觉。不少故事和过程，特别是一些重要的技术决策，我都是第一次知道完

整的来龙去脉。其中，谷歌的“三驾马车”、SQL 与 NoSQL 之争、从 MongoDB 到 CosmosDB 都是我们亲身经历的事件。我对微软相关的文章尤其感兴趣，因为其中的商业和技术决策是很好的学习案例，值得再三回顾：从 Cosmos 到 CosmosDB，从闭源到开源，从“面子”到“里子”，其中的兴衰一言难尽，值得所有技术公司借鉴。书中还分析了一些很新的系统，比如阿里巴巴的 Blink 系统等。

如果你对计算机和互联网行业感兴趣，想了解一些伟大的技术和公司的发展史，那么这本书正适合你，它会告诉你很多有趣的故事。更重要的是，你可以通过这些故事来了解背后的原因，所谓知其然且知其所以然。

汪军华

微软杰出工程师（Distinguished Engineer）

| 推荐序二

极客时间 App 上的专栏“技术与商业案例解读”上线后不久，我们请专栏作者徐飞来公司做直播，给用户分享他写这个专栏背后的故事，于是我第一次见到了飞总。和任何一个技术出身的小伙伴一样，飞总有点羞涩，但和很多技术人又不太一样的是，飞总很有自己的观点，而且敢说。时间过去很久了，我还能很清楚地记得他指点江山般地分析一家当红公司的种种优势与劣势，畅快淋漓而有见地。

文如其人，飞总的专栏也是如此。每次看或者听专栏的内容时，我都会纳闷：这哥们儿的知识怎么那么渊博？好像没有他不了解的公司，没有他不知道的奇闻逸事。也许正是这个原因，在极客时间 App 一开始用户量还不大的情况下，飞总的专栏就有几千人订阅。因为工作原因，我个人特别喜欢了解那些科技公司的历史。在我从事的软件开发领域，技术更迭实在太快，如果不能从源头理清这些技术的发展脉络，就很容易迷失在历史的谜团里。

飞总的文章不仅有趣，而且很有参考价值。有一次，飞总分析了

当时很火热的大数据公司 Cloudera 的业务模式、竞争对手及财务状况，然后他断言 Cloudera 当时的估值虚高，大概是其实际价值的两倍。这篇文章发布后的第三天，就传出了 Cloudera 公司自砍一半估值上市的新闻。听起来有点不可思议，但这也从侧面说明飞总对大数据行业的了解有多么深刻。

本书是将飞总的专栏中大数据方面的内容抽出来整理而成的。里面既有对国内外当红的大数据产品的解析，比如 TiDB、麒麟（Kyligence）以及被阿里巴巴以 9000 万欧元收购的 Flink，又有国际知名公司的大数据发展史，公司包括阿里巴巴、微软、亚马逊、IBM 等，读起来都非常有趣。技术变革从来都不是一件容易的事情，但是不管技术怎么变动，每家公司都希望在保持业务平稳增长的状态下发展，在移动互联网时代下做到这一点很难，不能掉以轻心。

古人说“以史为镜，可以知兴替”。相信有心的读者朋友——不论是程序员还是在 IT 企业家，都会从本书受益良多。

霍泰稳

极客邦科技创始人兼 CEO

| 推荐序三

从 20 世纪 60 年代至今，信息技术的版图不断扩大，从大型机、小型机、PC，到操作系统、网络、数据库、ITOM、云计算，而大数据无疑是近十年来最引人瞩目的新领域之一。

与其他技术领域相比，大数据领域的故事尤其难讲，一是因为领域新，在一个领域稳定下来之前，是不容易讲清楚其来龙去脉的。二是大数据领域的技术性很强，涉及的细分领域也很多，外行很难理清技术和需求之间的关系，自然更难看清兴衰背后的深层根源。第三，大数据领域中的技术发明者和商业角逐者既有大数据技术的使用者（比如雅虎、Facebook），又有产品和技术提供者（比如 MongoDB、Hadoop、谷歌），更有在技术使用者和商品提供者之间转换的云厂商巨头（比如亚马逊、阿里巴巴、微软等），不熟悉该领域的专家很难看清楚其中的演进规律。第四，在商业模式上，大数据领域是少有的甚至可以说是第一个以开源模式为主导的技术领域。

飞总在大数据领域有扎实的学术功底和丰富的企业实践经验，而且他对大数据领域的企业发展历程也有浓厚的兴趣，进行了深入的思考，因此才为我们呈现了这本精彩的关于大数据技术的“战国志”。

书中讲述的大数据故事，主要围绕的是企业软件市场。国内的企业软件公司主要偏重于业务应用侧，而偏重于基础架构和工具领域的大多数公司才刚刚起步，成功经验很少，还需要经历一个漫长的摸索和学习过程。

回到几十年前，中国各行各业都和国外有巨大的差距，但靠着艰苦卓绝的努力，许多行业已经大大缩小甚至消除了和国外的差距。我们相信在企业软件领域也会这样。只要投身于企业软件领域的中国创业者们认认真真、踏踏实实地向成功者学习，遵循科学的方法，假以时日，一定会出现一大批企业软件领域的优秀企业。

相信飞总的这本书会引发各位软件业同仁的思考，投石水中，必有涟漪，激荡反复，必将汇入中国企业软件的大潮，汹涌澎湃，席卷未来。

陈傲寒

优锆科技联合创始人兼 CEO，
BMC、Opsware 中国区前总经理

| 推荐序四

从 2000 年前的数据仓库、BI 到现在的大数据、人工智能，从 Lambda 到 IOTA，从“数据湖”到“数据河”，数据行业的新概念和产品层出不穷。飞总在本书中把主要的大数据组件的缘起、成长和未来都讲述了一遍，堪称“大数据领域的现代简史”。

我们可能知道 Hadoop 的 Cloudera 和 Hortonworks 版本，却不知它们为何从雅虎的团队分离出来了；我们可能知道 Spark 是现在最流行的大数据处理框架，却不知道它的诞生缘于一场针对 MapReduce 的论战，更不知道学术界以中立著称的 ACM 参与了这场“大战”；我们知道 IBM 在云化和大数据的浪潮中没有跟上潮流，却不知道 IBM 的云化策略要比谷歌、AWS 早很多；我们在非常开心地使用着各种各样的开源组件，却不知道这些开源厂商都在艰难度日，而一味从开源社区索取利益的云厂商们却赚得盆满钵满。

不算厚的一本书却对近 20 年的大数据组件的兴衰做了小结，以古鉴今，能给我们这些大数据从业者带来很多思考：大数据行业最终

的形态将会是怎样的？开源看上去很美好，但是怎样做才可以保持良性循环？大公司应该如何拓展领域并实现创新？如果像雅虎、IBM 这样的“巨无霸”也难逃“创新者的窘境”，那么我们又该如何走下去？

相对于 IT 行业整体的发展，数据行业其实还处在一个很稚嫩的早期阶段，很难脱离业务独立存在的特性迫使数据行业不断推陈出新，社区持续活跃。而相对于美国已经取得绝对优势的芯片行业，国内的企业更容易在大数据领域找到并把握领先的机会，因为中国的人口和数据量有天然的优势。纵观全书，不难发现，与其说大数据技术是被创造出来的，不如说大数据技术是被庞大的数据量倒逼出来的。所以，我非常看好中国大数据领域的发展前景，我们孕育着下一代全球级别的大数据开源架构，究竟哪家公司可以脱颖而出呢？让我们拭目以待。

最后，再次推荐大数据行业的从业者及有志于打造下一代大数据引擎的创新者阅读此书，从 30 多个大数据开源组件的兴衰历史里汲取经验，少走一些弯路，让中国在大数据创新的赛道上领跑世界！

郭炜

易观 CTO，中国软件行业协会智能应用服务分会副主任委员，

全球中小企业创业联合会（ICSB）副会长，

鲲鹏会（TGO）北京董事会会长

| 推荐序五

有人的地方就有江湖，有江湖的地方就有故事。技术界的故事同样精彩纷呈。

我在数据分析、大数据行业从事了十几年的工作，虽然对技术、架构及一些公司的发展有一定了解，但直到读了飞总的这本书，才发现背后还有那么多的故事和纠缠，而且那么精彩和有趣。

飞总在书中如数家珍般地讲述了很多不为人知的故事，从数据库领域顶级科研会议、论文、流派，到商业模式以及相关技术方向和架构。飞总在数据库理论和实践领域有多年的积淀，并且拥有 VLDB、SIGMOD 论文评审的深厚功底，因此才能抽丝剥茧，为我们呈现一个个精彩的企业故事，勾勒出技术圈的一段至关重要的历史。

数据分析作为当今社会科技的基础技术，已经深入人们生活和工作的方方面面，从最早的信息化到数据仓库、大数据、云端数据分析，再到人工智能技术的初期应用，到处体现了数据、数据分析方法和技术的重要性。毫不夸张地说，没有数据分析，今天的大部分业务和工