高等职业教育大数据工程技术系列教材

大数据平台运维基础

龚大丰　翁正秋　池万乐　主　编

施莉莉　王小铭　副主编

電子工業出版社

Publishing House of Electronics Industry

北京・BEIJING

内 容 简 介

本书是高等职业教育大数据技术与应用系列教材中的一册，讲解了大数据系统运行维护过程中的各个主要任务，包括大数据生态圈、Hadoop 环境搭建与运维、Hive 环境搭建与基本操作、HBase 环境搭建与运维、Hadoop 常用组件安装等内容。本书内容详尽充实，针对每个知识点都配有相应的实验用于验证和巩固，在基础理论知识上增加了运维大数据平台实践应用知识，重点介绍了大数据系统的运维实操技能，对于培养应用型大数据平台运维人才有着很强的指导性。

本书既可以作为高等职业院校大数据平台运维课程的教学用书，也同样适合作为有志从事大数据系统运维工作的广大爱好者的参考书。

图书在版编目（CIP）数据

大数据平台运维基础 / 龚大丰，翁正秋，池万乐主编. —北京：电子工业出版社，2022.6
ISBN 978-7-121-43420-4

Ⅰ. ①大… Ⅱ. ①龚… ②翁… ③池… Ⅲ. ①数据处理－高等职业教育－教材 Ⅳ. ①TP274

中国版本图书馆 CIP 数据核字（2022）第 077348 号

责任编辑：徐建军　　文字编辑：徐　萍
印　　刷：
装　　订：
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：787×1 092　1/16　印张：13.5　字数：346 千字
版　　次：2022 年 6 月第 1 版
印　　次：2022 年 6 月第 1 次印刷
印　　数：1 200 册　　定价：46.00 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 zlts@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）88254570，xujj@phei.com.cn。

前言
Preface

今天，越来越多的行业对大数据应用表现出强烈的兴趣。大数据或者相关数据分析解决方案的使用不但出现在互联网行业，像电信、金融、能源这些传统行业，越来越多的用户也开始尝试使用大数据解决具体业务问题，来提升自己的业务水平。在“大数据”背景之下，精通“大数据”的专业人才将成为企业重要的业务角色，“大数据”从业人员薪酬持续增长，人才缺口巨大。

大数据运维工程师作为大数据专业培养的基础岗位，在国民经济的各个领域都有很大的需求，基本上哪里有大数据哪里就需要大数据运维工程师。大数据运维工程师的工作内容包括：大数据集群的运维工作（Hadoop、HBase、Hive 等）；负责大数据集群性能优化、扩容；负责 Hadoop 集群的监控、数据备份、数据监控、报警、故障处理；研究大数据运维相关技术，根据系统需求制定运维技术方案，开发自动化运维工具和运维辅助系统；研究大数据业务相关运维技术，优化集群服务架构，探索新的大数据运维技术及发展方向。

本书作为培养应用型大数据运维工程师的基础教材，覆盖了大数据运维工作的各个方面，在基础理论知识上增加了运维大数据平台实践应用知识，重点介绍了大数据系统的运维实操技能，既适合大数据运维工程师学习使用，也可作为已经从事大数据运维工作人员的参考书。

本书由温州职业技术学院大数据技术专业国家级职业教育教师教学创新团队与章鱼大数据（优选创新（北京）科技有限公司）组织策划，由龚大丰、翁正秋、池万乐担任主编，施莉莉、王小铭担任副主编。其中，第 1、3 章由池万乐编写，第 2 章由翁正秋编写，第 4、5 章由龚大丰编写，实验部分由施莉莉和王小铭参与编写，全书由龚大丰统稿。此外，参与编写工作的还有陈贤、邵剑集、高瑜澧、陈清华、施郁文、杜益虹等。同时，也特别感谢温州市大数据发展管理局陈力琼为本书提供了修订意见。

本书的编写得到温州职业技术学院教改项目（项目编号：WZYYFFP2020005、WZYSZZY2104、WZYSZKC2106、WZYzd202003、WZYCJRH201905、WZYZD201810）以及浙江省产学合作协同育人项目（“基于政产学研用的信息技术类专业课证融通改革”浙教办函〔2021〕7 号）立项支持，在此表示衷心的感谢。

为了方便教师教学，本书配有电子教学课件及相关资源，请有此需要的教师登录华信教育资源网（www.hxedu.com.cn）注册后免费进行下载，如有问题可在网站留言板留言或与电子工业出版社联系（E-mail：hxedu@phei.com.cn）。

教材建设是一项系统工程，需要在实践中不断加以完善及改进，同时由于时间仓促、编者水平有限，书中难免存在疏漏和不足之处，敬请同行专家与广大读者批评指正。

编　者

目录
Contents

第1章 大数据生态圈

学习任务

对大数据在概念上有一个宏观的认识，同时了解大数据的发展与主要技术。

- ☑ 了解大数据的概念和价值。
- ☑ 了解大数据的特点。
- ☑ 了解大数据技术组成与生态圈。
- ☑ 了解大数据的行业应用和未来发展。

知识点

- ☑ 大数据的概念介绍。
- ☑ 大数据的特点介绍。
- ☑ 大数据的主要技术介绍。
- ☑ 大数据的应用介绍。

1.1 大数据的概念和价值

很多人对于这些热门的新技术、新趋势往往趋之若鹜却又很难说得透彻，如果问他大数据和其有什么关系，估计很少能说出个一二三来。究其原因，一是因为大家对新技术有着相同的原始渴求，至少知其然在聊天时不会显得很“落伍”；二是在工作和生活环境中真正能参与实践大数据的案例实在太少了，所以大家没有必要花时间去知其所以然。

如果说大数据就是数据大，或者侃侃而谈 4 个“V”，也许很有深度地谈到 BI 或预测的价值，又或者拿 Google 和 Amazon 举例，技术流可能会聊起 Hadoop 和 CloudComputing，不管对错，都无法描绘出对大数据的整体认识，不说是片面，但至少有些管窥蠡测、隔靴瘙痒了。

大数据就是互联网发展到现今阶段的一种表象或特征而已，没有必要神话它或将其视为深

不可测。在以云计算为代表的技术创新大幕的衬托下，这些原本很难收集和使用的数据开始容易地被利用起来了，通过各行各业的不断创新，大数据会逐步为人类创造更多的价值。

1. 大数据的概念

最早提出大数据时代到来的是麦肯锡："数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。人们对于海量数据的挖掘和运用，预示着新一波生产率增长和消费者盈余浪潮的到来。"

业界（IBM 最早定义）将大数据的特征归纳为 4 个"V"（量，Volume；多样，Variety；价值，Value；速度，Velocity），或者说特点有四个层面：第一，数据体量巨大，大数据的起始计量单位至少是 P（1000T）、E（100 万 T）或 Z（10 亿 T）；第二，数据类型繁多，比如，网络日志、视频、图片、地理位置信息等；第三，价值密度低，商业价值高；第四，处理速度快。最后这一点也是和传统的数据挖掘技术有着本质的不同。

俗话说：三分技术，七分数据，得数据者得天下。先不论谁说的，但是这句话的正确性已经不用去论证了。维克托·迈尔-舍恩伯格在《大数据时代》一书中列举了诸多例证，都是为了说明一个道理：在大数据时代已经到来的时候要用大数据思维去发掘大数据的潜在价值。书中，作者提及最多的是 Google 如何利用人们的搜索记录挖掘数据二次利用价值，比如预测某地流感爆发的趋势；Amazon 如何利用用户的购买和浏览历史数据进行有针对性的书籍购买推荐，以此有效提升销售量；Farecast 如何利用过去十年所有的航线机票价格打折数据，来预测用户购买机票的时机是否合适。

那么，什么是大数据思维？维克托·迈尔-舍恩伯格认为：①需要全部数据样本而不是抽样；②关注效率而不是精确度；③关注相关性而不是因果关系。

阿里巴巴的王坚对于大数据也有一些独特的见解，比如：

"今天的数据不是大，真正有意思的是数据变得在线了，这个恰恰是互联网的特点。"

"非互联网时期的产品，功能一定是它的价值，今天互联网的产品，数据一定是它的价值。"

"你千万不要想着拿数据去改进一个业务，这不是大数据。你一定是去做了一件以前做不了的事情。"

特别是最后一点，笔者是非常认同的，大数据的真正价值在于创造，在于填补无数个还未实现过的空白。

有人把数据比喻为蕴藏能量的煤矿。煤炭按照性质有焦煤、无烟煤、肥煤、贫煤等分类，而露天煤矿、深山煤矿的挖掘成本又不一样。与此类似，大数据并不在于"大"，而在于"有用"。价值含量、挖掘成本比数量更为重要。

2. 大数据的价值

大数据是什么？在投资者眼里是金光闪闪的两个字：资产。比如，Facebook 上市时，评估机构评定的有效资产中大部分都是其社交网站上的数据。

如果把大数据比作一种产业，那么这种产业实现盈利的关键，在于提高对数据的"加工能力"，通过"加工"实现数据的"增值"。

Target 超市以 20 多种怀孕期间孕妇可能会购买的商品为基础，将所有用户的购买记录作为数据来源，通过构建模型分析购买者的行为相关性，能准确地推断出孕妇的具体临盆时间，这样 Target 的销售部门就可以有针对地在每个怀孕顾客的不同阶段寄送相应的产品优惠券。

Target 的例子是一个很典型的案例，这就印证了维克托·迈尔-舍恩伯格提过的一个很有指导意义的观点：通过找出一个关联物并监控它，就可以预测未来。Target 通过监测购买者购买商品的时间和品种来准确预测顾客的孕期，这就是对数据的二次利用的典型案例。比如，我们通过采集驾驶员手机的 GPS 数据，就可以分析出当前哪些道路正在堵车，并可以及时发布道路交通提醒；通过采集汽车的 GPS 位置数据，就可以分析城市的哪些区域停车较多，这也代表该区域有着较为活跃的人群。

不管大数据的核心价值是不是预测，但是基于大数据形成决策的模式已经为不少的企业带来了盈利和声誉。

从大数据的价值链条来分析，存在三种模式：

第一，手握大数据，但是没有利用好。比较典型的是金融机构、电信行业、政府机构等。

第二，没有数据，但是知道如何帮助有数据的人利用它。比较典型的是 IT 咨询和服务企业，如埃森哲、IBM、Oracle 等。

第三，既有数据，又有大数据思维。比较典型的是 Google、Amazon、Mastercard 等。

未来在大数据领域最具有价值的是两种事物：①拥有大数据思维的人，这种人可以将大数据的潜在价值转化为实际利益；②还未被大数据触及过的业务领域。这些是还未被挖掘的油井、金矿，是所谓的蓝海。

Walmart 作为零售行业的巨头，其分析人员会对每个阶段的销售记录进行全面的分析。有一次他们无意中发现了虽不相关但很有价值的数据，在美国的飓风来临季节，超市的蛋挞和抵御飓风物品竟然销量都有大幅增加，于是他们做了一个明智的决策，就是将蛋挞的销售位置移到了飓风物品销售区域旁边，看起来是为了方便用户挑选，但是没有想到蛋挞的销量因此又提高了很多。

还有一个有趣的例子，1948 年辽沈战役期间，司令员林彪要求每天要进行例常的“每日军情汇报”，由值班参谋读出下属各个纵队、师、团用电台报告的当日战况和缴获情况。那几乎是重复着千篇一律枯燥无味的数据：每支部队歼敌多少、俘虏多少；缴获的火炮、车辆多少，枪支、物资多少……有一天，参谋照例汇报当日的战况，林彪突然打断他：“刚才念的在胡家窝棚那个战斗的缴获，你们听到了吗？”大家都很茫然，因为如此战斗每天都有几十起，不都是差不多一模一样的枯燥数字吗？林彪扫视一周，见无人回答，便接连问了三句：“为什么那里缴获的短枪与长枪的比例比其他战斗略高？”“为什么那里缴获和击毁的小车与大车的比例比其他战斗略高？”“为什么在那里俘虏和击毙的军官与士兵的比例比其他战斗略高？”林彪司令员大步走向挂满军用地图的墙壁，指着地图上的那个点说：“我猜想，不，我断定！敌人的指挥所就在这里！”果然，部队很快就抓住了敌方的指挥官廖耀湘，并取得这场重要战役的胜利。

这些例子真实地反映在各行各业，探求数据价值取决于把握数据的人，关键是人的数据思维；与其说是大数据创造了价值，不如说是大数据思维触发了新的价值增长。

1.2　大数据的特点

大数据是一个较为抽象的概念，正如信息学领域大多数新兴概念一样，大数据至今尚无确切、统一的定义。

在维基百科中关于大数据的定义为：大数据是指利用常用软件工具来获取、管理和处理数据所耗时间超过可容忍时间的数据集。这并不是一个精确的定义，因为无法确定常用软件工具的范围，可容忍时间也是个概略的描述。IDC 对大数据做出的定义为：大数据一般会涉及两种或两种以上数据形式。它要收集超过 100TB 的数据，并且是高速、实时数据流；或者是从小数据开始，但数据每年会增长 60%以上。这个定义给出了量化标准，但只强调数据量大、种类多、增长快等数据本身的特征。

研究机构 Gartner 给出了这样的定义：大数据是需要新处理模式才能具有更强的决策力、洞察发现力和流程优化能力的海量、高增长率和多样化的信息资产。这也是一个描述性的定义，在对数据描述的基础上加入了处理此类数据的一些特征，用这些特征来描述大数据。

当前，较为统一的认识是大数据有四个基本特点：数据量大（Volume）、数据类型多样（Variety）、数据产生和处理速度快（Velocity）、数据价值密度低（Value）。再加上数据真实性（Veracity），构成所谓的五“V”特性。这些特性使得大数据有别于传统的数据概念。

大数据的概念与“海量数据”不同，后者只强调数据的量，而大数据不仅用来描述大量的数据，还进一步指出数据的复杂形式、数据的快速时间特性，以及对数据的分析、处理等专业化处理，最终获得有价值信息的能力。

1. 数据量大（Volume）

大数据聚合在一起的数据量是非常大的，根据 IDC 的定义至少要有超过 100TB 的可供分析的数据，数据量大是大数据的基本属性。导致数据规模激增的原因有很多，首先是随着互联网络的广泛应用，使用网络的人、企业、机构增多，数据获取、分享变得相对容易。以前，只有少量的机构可以通过调查、取样的方法获取数据，同时发布数据的机构也很有限，人们难以在短期内获取大量的数据，而现在用户可以通过网络非常方便地获取数据，同时用户在有意的分享和无意的单击、浏览中都可以快速地提供大量数据。其次是随着各种传感器数据获取能力的大幅提高，使得人们获取的数据越来越接近原始事物本身，描述同一事物的数据量激增。早期的单位化数据，对原始事物进行了一定程度的抽象，数据维度低，数据类型简单，多采用表格的形式来收集、存储、整理，数据的单位、量纲和意义基本统一，存储、处理的只是数值而已，因此数据量有限，增长速度慢。而随着应用的发展，数据维度越来越高，描述相同事物所需的数据量越来越大。以当前最为普遍的网络数据为例，早期网络上的数据以文本和一维的音频为主，维度低，单位数据量小。近年来，图像、视频等二维数据大规模涌现，随着三维扫描设备及 Kinect 等动作捕捉设备的普及，数据越来越接近真实的世界，数据的描述能力不断增强，而数据量本身必将以几何级数增长。此外，数据量大还体现在人们处理数据的方法和理念发生了根本的改变。早期，人们对事物的认知受限于获取、分析数据的能力，一直利用采样的方法，以少量的数据来近似地描述事物的全貌，样本的数量可以根据数据获取、处理能力来设定。不管事物多么复杂，通过采样得到部分样本，数据规模变小，就可以利用当时的技术手段来进行数据管理和分析，如何通过正确的采样方法以最小的数据量尽可能分析整体属性成了当时的重要问题。随着技术的发展，样本数目逐渐逼近原始的总体数据，且在某些特定的应用领域，采样数据可能远不能描述整个事物，可能丢掉大量重要细节，甚至可能得到完全相反的结论，因此，当今有直接处理所有数据而不是只考虑采样数据的趋势。使用所有的数据可以带来更高的精确性，从更多的细节来解释事物属性，同时必然使得要处理的数据量显著增多。

2. 数据类型多样（Variety）

数据类型繁多、复杂多变是大数据的重要特性。以往的数据尽管数量庞大，但通常是事先定义好的结构化数据。结构化数据是将事物向便于人类和计算机存储、处理、查询的方向抽象的结果，结构化在抽象的过程中，忽略了一些在特定的应用下可以不考虑的细节，抽取了有用的信息。处理此类结构化数据，只需事先分析好数据的意义及数据间的相关属性，构造表结构来表示数据的属性，数据都以表格的形式保存在数据库中，数据格式统一，以后不管再产生多少数据，只需根据其属性，将数据存储在合适的位置，就可以方便地处理、查询，一般不需要为新增的数据显著地更改数据聚集、处理、查询方法，限制数据处理能力的只是运算速度和存储空间。这种关注结构化信息，强调大众化、标准化的属性使得处理传统数据的复杂程度一般呈线性增长，新增的数据可以通过常规的技术手段处理。随着互联网络与传感器的飞速发展，非结构化数据大量涌现，非结构化数据没有统一的结构属性，难以用表结构来表示，在记录数据数值的同时还需要存储数据的结构，增加了数据存储、处理的难度。而时下在网络上流动着的数据大部分是非结构化数据，人们上网不只是看看新闻、发送文字邮件，还会上传/下载照片、视频、发送微博等非结构化数据，同时，遍及工作、生活中各个角落的传感器也不断地产生各种半结构化、非结构化数据，这些结构复杂、种类多样，同时规模又很大的半结构化、非结构化数据逐渐成为主流数据。如上所述，非结构化数据量已占到数据总量的 75%以上，且非结构化数据的增长速度比结构化数据快 10～50 倍。在数据激增的同时，新的数据类型层出不穷，已经很难用一种或几种规定的模式来表征日趋复杂、多样的数据形式，这样的数据已经不能用传统的数据库表格来整齐地排列、表示。大数据正是在这样的背景下产生的，大数据与传统数据处理最大的不同就是重点关注非结构化信息，大数据关注包含大量细节信息的非结构化数据，强调小众化、体验化的特性使得传统的数据处理方式面临巨大的挑战。

3. 数据产生和处理速度快（Velocity）

要求数据的快速处理，是大数据区别于传统海量数据处理的重要特性之一。随着各种传感器和互联网络等信息获取、传播技术的飞速发展普及，数据的产生、发布越来越容易，产生数据的途径增多，个人甚至成为数据产生的主体之一，数据呈爆炸式快速增长，新数据不断涌现，快速增长的数据量要求数据处理的速度也要相应地提升，才能使大量的数据得到有效的利用，否则不断激增的数据不但不能为解决问题带来优势，反而成了快速解决问题的负担。同时，数据不是静止不动的，而是在互联网络中不断流动，且通常这样的数据其价值是随着时间的推移而迅速降低的，如果数据尚未得到有效的处理，就失去了价值，大量的数据就没有意义了。此外，在许多应用中要求能够实时处理新增的大量数据，比如，有大量在线交互的电子商务应用，就具有很强的时效性，大数据以数据流的形式产生、快速流动、迅速消失，且数据流量通常不是平稳的，会在某些特定的时段激增，数据的涌现特征明显。而用户对于数据的响应时间通常非常敏感，心理学实验证实，从用户体验的角度来看，瞬间（Moment，3s）是可以容忍的最大极限。对于大数据应用而言，很多情况下都必须在 1s 或者瞬间内形成结果，否则处理结果就是过时和无效的，这种情况下，大数据要求快速、持续的实时处理。对不断激增的海量数据的实时处理要求，是大数据与传统海量数据处理技术的关键差别之一。

4. 数据价值密度低（Value）

数据价值密度低是大数据关注的非结构化数据的重要属性。传统的结构化数据，依据特定的应用，对事物进行了相应的抽象，每一条数据都包含该应用需要考量的信息。而大数据为了获取事物的全部细节，不对事物进行抽象、归纳等处理，直接采用原始的数据，保留了数据的原貌，且通常不对数据进行采样，直接采用全体数据。由于减少了采样和抽象，呈现所有数据和全部细节信息，可以分析更多的信息，但也引入了大量没有意义的信息，甚至是错误的信息，因此相对于特定的应用，大数据关注的非结构化数据的价值密度偏低。以当前广泛应用的监控视频为例，在连续的监控过程中，大量的视频数据被存储下来，许多数据可能是无用的，对于某一特定的应用，比如获取犯罪嫌疑人的体貌特征，有效的视频数据可能仅仅有一两秒，大量不相关的视频信息增加了获取这有效的一两秒数据的难度。但是大数据的数据密度低是指相对于特定的应用，有效的信息相对于数据整体是偏少的；信息有效与否也是相对的，对于某些应用是无效的信息可能对于另外一些应用却成为最关键的信息；数据的价值也是相对的，有时一条微不足道的细节数据可能造成巨大的影响，比如网络中的一条几十个字符的微博，就可能通过转发而快速扩散，导致相关的信息大量涌现，其价值不可估量。因此，为了保证对于新产生的应用有足够的有效信息，通常必须保存所有数据，这样就使得一方面是数据的绝对数量激增；另一方面是数据包含有效信息量的比例不断减小，数据价值密度偏低。

5. 数据真实性（Veracity）

数据真实性即数据的准确性和可信赖度，或者叫数据的质量。大数据中的内容是与真实世界中事的发生息息相关的，研究大数据就是从庞大的网络数据中提取出能够解释和预测现实事件的过程。

1.3 大数据技术组成与生态圈

1. 云技术

大数据常和云计算联系到一起，因为实时的大型数据集分析需要分布式处理框架来向数十台、数百台甚至数万台的计算机分配工作。可以说，云计算充当了工业革命时期发动机的角色，而大数据则是电。

云计算思想的起源是麦卡锡在 20 世纪 60 年代提出的：把计算能力作为一种像水和电一样的公用事业提供给用户。

如今，在 Google、Amazon、Facebook 等一批互联网企业的引领下，一种行之有效的模式出现了：云计算提供基础架构平台，大数据应用运行在这个平台上。

业内是这么形容两者的关系的：没有大数据的信息积淀，则云计算的计算能力再强大，也难以找到用武之地；没有云计算的处理能力，则大数据的信息积淀再丰富，也终究只是镜花水月。

那么大数据到底需要哪些云计算技术呢？

这里暂且列举一些，比如虚拟化技术、分布式处理技术、海量数据的存储和管理技术、

NoSQL、实时流数据处理、智能分析技术（类似模式识别及自然语言理解）等。

云计算和大数据两者结合后会产生如下效应：可以提供更多基于海量业务数据的创新型服务；通过云计算技术的不断发展降低大数据业务的创新成本。

如果将云计算与大数据进行一些比较，最明显的区别在两个方面：

第一，在概念上两者有所不同，云计算改变了 IT，而大数据则改变了业务。然而大数据必须有云作为基础架构，才能得以顺畅运营。

第二，大数据和云计算的目标受众不同，云计算是 CIO 等关心的技术层，是一个进阶的 IT 解决方案，而大数据是 CEO 关注的，是业务层的产品，大数据的决策者是业务层。

2. 分布式处理技术

分布式处理系统可以将不同地点或具有不同功能或拥有不同数据的多台计算机用通信网络连接起来，在控制系统的统一管理控制下，协调地完成信息处理任务。

以 Hadoop（Yahoo）为例进行说明，Hadoop 是一个实现了 MapReduce 模式的能够对大量数据进行分布式处理的软件框架，是以一种可靠、高效、可伸缩的方式进行数据处理的。

而 MapReduce 是 Google 提出的一种云计算的核心计算模式，是一种分布式运算技术，也是简化的分布式编程模式。MapReduce 模式的主要思想是通过将要执行的任务（如程序）拆解成 Map（映射）和 Reduce（化简）的方式，在数据被分割后通过 Map 函数的程序将数据映射成不同的区块，分配给计算机机群处理、达到分布式运算的效果，再通过 Reduce 函数的程序将结果汇整，从而输出开发者需要的结果。

再来看看 Hadoop 的特性，首先，它是可靠的，因为它假设计算元素和存储会失败，因此它维护多个工作数据副本，确保能够针对失败的节点重新分布处理；其次，Hadoop 是高效的，因为它以并行的方式工作，通过并行处理加快处理速度；再次，Hadoop 还是可伸缩的，能够处理 PB 级数据，最后，Hadoop 依赖于社区服务器，因此它的成本比较低，任何人都可以使用。

也可以这么理解 Hadoop 的构成：Hadoop=HDFS（文件系统，数据存储技术相关）+HBase（数据库）+MapReduce（数据处理）+……（其他）。

Hadoop 用到的一些技术如下。

- HDFS：Hadoop 分布式文件系统（Distributed File System）——HDFS（Hadoop Distributed File System）。
- MapReduce：并行计算框架。
- HBase：类似 Google BigTable 的分布式 NoSQL 系列数据库。
- Hive：数据仓库工具，由 Facebook 贡献。
- ZooKeeper：分布式锁设施，提供类似 Google Chubby 的功能，由 Facebook 贡献。
- Avro：新的数据序列化格式与传输工具，将逐步取代 Hadoop 原有的 IPC 机制。
- Pig：大数据分析平台，为用户提供多种接口。
- Ambari：Hadoop 管理工具，可以快捷地监控、部署、管理集群。
- Sqoop：用于在 Hadoop 与传统的数据库之间进行数据的传递。

说了这么多，举一个实际的例子，虽然这个例子有些陈旧。淘宝的海量数据技术架构有助于我们理解大数据的运作处理机制，如图 1.1 所示。

淘宝的海量数据产品技术架构分为五个层次，从上至下分别是数据源、计算层、存储层、查询层和产品层。

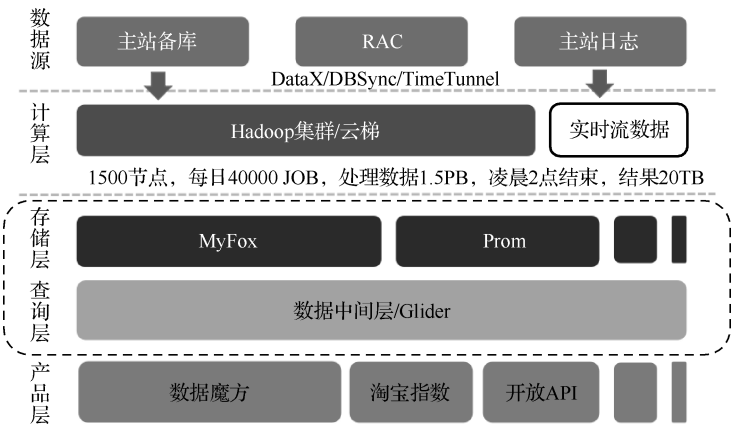

图 1.1　淘宝海量数据技术架构

- 数据来源层：存放着淘宝各店的交易数据。在该层产生的数据，通过 DataX、DBSync 和 TimeTunnel 实时地传输到下面即将介绍的“云梯”。
- 计算层：在计算层内，淘宝采用的是 Hadoop 集群，这个集群，我们暂且称为云梯，是计算层的主要组成部分。在云梯上，系统每天会对数据产品进行不同的 MapReduce 计算。
- 存储层：在这一层，淘宝采用了两个东西，一个是 MyFox，一个是 Prom。MyFox 是基于 MySQL 的分布式关系型数据库的集群，Prom 是基于 Hadoop Hbase 技术的一个 NoSQL 的存储集群。
- 查询层：在这一层中，Glider 是以 HTTP 协议对外提供 restful 方式的接口。数据产品通过一个唯一的 URL 来获取它想要的数据。同时，数据查询也是通过 MyFox 来完成的。
- 产品层：这是最后一层，这个就不用解释了。

3. 存储技术

大数据可以抽象地分为大数据存储和大数据分析，这两者的关系是：大数据存储的目的是支撑大数据分析。到目前为止，它们还是两种截然不同的计算机技术领域：大数据存储致力于研发可以扩展至 PB 级别甚至 EB 级别的数据存储平台；大数据分析关注在最短时间内处理大量不同类型的数据集。

提到存储，有一个著名的摩尔定律相信大家都听过：每过 18 个月集成电路的复杂性就增加一倍。所以，存储器的成本每 18～24 个月就下降一半。成本的不断下降也造就了大数据的可存储性。

比如，Google 大约管理着超过 50 万台服务器和 100 万块硬盘，而且 Google 还在不断地扩大计算能力和存储能力，其中很多的扩展都是在廉价服务器和普通存储硬盘的基础上进行的，这大大降低了其服务成本，因此可以将更多的资金投入技术的研发当中。

以 Amazon 举例，Amazon S3 是一种面向 Internet 的存储服务。该服务旨在让开发人员能更轻松地进行网络规模计算。Amazon S3 提供一个简明的 Web 服务界面，用户可通过它随时在 Web 上的任何位置存储和检索任意大小的数据。此服务让所有开发人员都能访问同一个具备高扩展性、可靠性、安全性和快速价廉的基础设施，Amazon 用它来运行其全球的网站网络。再看看 S3 的设计指标：在特定年度内为数据元提供 99.999999999%的耐久性和 99.99%的可用性，并能够承受两个设施中的数据同时丢失。

S3 很成功也确实卓有成效，S3 云的存储对象已达到万亿级别，而且性能表现相当良好。S3 云已经拥有万亿跨地域存储对象，同时 AWS 的对象执行请求也达到百万的峰值数量。目前全球范围内已经有数以十万计的企业在通过 AWS 运行自己的全部或者部分日常业务。这些企业用户遍布 190 多个国家，几乎世界上的每个角落都有 Amazon 用户的身影。

4. 感知技术

大数据的采集和感知技术的发展是紧密联系的。以传感器技术、指纹识别技术、RFID 技术、坐标定位技术等为基础的感知能力提升同样是物联网发展的基石。全世界的工业设备、汽车、电表上有着无数的数码传感器，随时测量和传递着有关位置、运动、震动、温度、湿度乃至空气中化学物质的变化等信息，产生海量的数据信息。

随着智能手机的普及，感知技术可谓迎来了发展的高峰期，除了地理位置信息被广泛地应用外，一些新的感知手段也开始登上舞台。比如光线传感器，类似于手机的眼睛。人类的眼睛能在不同光线的环境下，调整进入眼睛的光线，光线传感器则可以让手机感测环境光线的强度，进而调节手机屏幕的亮度。运用光线传感器来协助调整屏幕亮度，能进一步达到延长电池寿命的作用。光线传感器也可搭配其他传感器一同来侦测手机是否被放置在口袋中，以防止误触。超声波指纹传感器不会受到汗水、油污的干扰，辨识速度也更快，运用在手机中可以完成解锁、加密、支付等。在一些户外应用中需要测量气压值时，搭配气压传感器的手机也能派上用场，在 iOS 的健康应用中，可以计算出一个人爬了几层楼。心率传感器通过高亮度的 LED 灯照射手指，因心脏将血液压送到毛细血管时，亮度（红光的深度）会呈现周期性的变化，使用摄影机捕捉这些规律性的变化，并将数据传送到手机中进行运算，进而判断心脏的收缩频率，就能得出每分钟的心跳数。

除此之外，还有很多与感知相关的技术革新让我们耳目一新：牙齿传感器实时监控口腔活动及饮食状况，婴儿穿戴设备可用大数据去养育宝宝，Intel 正研发 3D 笔记本摄像头可追踪眼球读懂情绪，日本公司开发新型可监控用户心率的纺织材料，业界正在尝试将生物测定技术引入支付领域等。

其实，这些感知被逐渐捕获的过程就是世界被数据化的过程。一旦世界被完全数据化了，那么世界的本质也就是信息了。

就像一句名言所说，“人类以前延续的是文明，现在传承的是信息”。

1.4　大数据的行业应用和未来发展

1. 大数据的行业应用

我们先看看大数据在当下有怎样的杰出表现：

大数据帮助政府实现市场经济调控、公共卫生安全防范、灾难预警、社会舆论监督；

大数据帮助城市预防犯罪，实现智慧交通，提升紧急应急能力；

大数据帮助医疗机构建立患者的疾病风险跟踪机制，帮助医药企业提升药品的临床使用效果，帮助艾滋病研究机构为患者提供定制的药物；

大数据帮助航空公司节省运营成本，帮助电信企业实现售后服务质量提升，帮助保险企业

识别欺诈骗保行为，帮助快递公司监测分析运输车辆的故障险情以提前预警维修，帮助电力公司有效识别预警找出即将发生故障的设备；

大数据帮助电商公司向用户推荐商品和服务，帮助旅游网站为旅游者提供心仪的旅游路线，帮助二手市场的买卖双方找到最合适的交易目标，帮助用户找到最合适的商品购买时期、商家和最优惠的价格；

大数据帮助企业提升营销的针对性，降低物流和库存的成本，减小投资的风险，帮助企业提升广告投放精准度；

大数据帮助娱乐行业预测歌手、歌曲、电影、电视剧的受欢迎程度，并为投资者分析评估拍一部电影需要投入多少钱才最合适，否则就有可能收不回成本；

大数据帮助社交网站提供更准确的好友推荐，为用户提供更精准的企业招聘信息，向用户推荐可能喜欢的游戏及适合购买的商品。

互联网上的数据每年增长 50%，每两年便翻一番，而目前世界上 90%以上的数据是最近几年才产生的。据 IDC 估计，2020 年全球总共拥有 35ZB 的数据量。互联网是大数据发展的前沿阵地，随着 Web 2.0 时代的发展，人们似乎已经习惯了将自己的生活通过网络进行数据化，方便分享、记录和回忆。

互联网上的大数据很难清晰地界定分类界限，我们先看看 BAT 的大数据。

百度拥有两种类型的大数据：用户搜索表征的需求数据；爬虫和阿拉丁获取的公共 Web 数据。搜索巨头百度围绕数据而生，它对网页数据的爬取、网页内容的组织和解析，通过语义分析对搜索需求的精准理解进而从海量数据中找准结果，以及精准的搜索引擎关键字广告，实质上就是一个数据的获取、组织、分析和挖掘的过程。搜索引擎在大数据时代面临的挑战有：更多的暗网数据；更多的 Web 化但是没有结构化的数据；更多的 Web 化、结构化但是封闭的数据。

阿里巴巴拥有交易数据和信用数据，这两种数据更容易变现，挖掘出商业价值。除此之外，阿里巴巴还通过投资等方式掌握了部分社交数据、移动数据，如微博和高德。

腾讯拥有用户关系数据和基于此产生的社交数据。这些数据可以分析人们的生活和行为，从其中挖掘出政治、社会、文化、商业、健康等领域的信息，甚至可以预测未来。

在信息技术更为发达的美国，除了行业知名的类似 Google、Facebook 等巨头外，还涌现了很多大数据类型的公司，它们专门经营数据产品。

- Metamarkets：该公司对 Twitter、支付、签到和一些与互联网相关的问题进行了分析，为客户提供了很好的数据分析支持。
- Tableau：该公司的精力主要集中于将海量数据以可视化的方式展现出来。其为数字媒体提供了一个新的展示数据的方式。他们提供一个免费工具，任何人在没有编程知识背景的情况下都能制作出数据专用图表。这个软件还能对数据进行分析，并提供有价值的建议。
- ParAccel：该公司向美国执法机构提供数据分析，比如对 15000 个有犯罪前科的人进行跟踪，从而向执法机构提供参考性较高的犯罪预测。
- QlikTech：QlikTech 旗下的 Qlikview 是一个商业智能领域的自主服务工具，能够应用于科学研究和艺术等领域。为了帮助开发者对这些数据进行分析，QlikTech 提供了具备对原始数据进行可视化处理等功能的工具。
- GoodData：GoodData 希望帮助客户从数据中挖掘财富。这家公司主要面向商业用户和

IT 企业高管，提供数据存储、性能报告、数据分析等工具。

- TellApart：TellApart 和电商公司进行合作，他们会根据用户的浏览行为等数据进行分析，通过锁定潜在买家的方式提高电商企业的收入。
- DataSift：DataSift 主要收集并分析社交网络媒体上的数据，帮助品牌公司掌握突发新闻的舆论点，并制订有针对性的营销方案。这家公司还和 Twitter 有合作协议，把自己变成了行业中为数不多的可以分析早期 Tweet 的创业公司。
- Datahero：该公司的目标是将复杂的数据变得更加简单明了，方便普通人去理解和想象。

举了很多例子，这里简要归纳一下，互联网中大数据的典型代表性包括：

用户行为数据（精准广告投放、内容推荐、行为习惯和喜好分析、产品优化等）；

用户消费数据（精准营销、信用记录分析、活动促销、理财等）；

用户地理位置数据（O2O 推广、商家推荐、交友推荐等）；

互联网金融数据（P2P、小额贷款、支付、信用、供应链金融等）；

用户社交等 UGC 数据（趋势分析、流行元素分析、受欢迎程度分析、舆论监控分析、社会问题分析等）。

之前，奥巴马政府宣布投资 2 亿美元拉动大数据相关产业发展，将“大数据战略”上升为国家意志。奥巴马政府将数据定义为“未来的新石油”，并表示一个国家拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，未来，对数据的占有和控制甚至将成为陆权、海权、空权之外的另一种国家核心资产。

在国内，政府各个部门都握有构成社会基础的原始数据，比如，气象数据、金融数据、信用数据、电力数据、煤气数据、自来水数据、道路交通数据、客运数据、安全刑事案件数据、住房数据、海关数据、出入境数据、旅游数据、医疗数据、教育数据、环保数据，等等。这些数据在每个政府部门里看起来是单一的、静态的。但是，如果政府可以将这些数据关联起来，并对这些数据进行有效的关联分析和统一管理，那么这些数据必将获得新生，其价值是无法估量的。

具体来说，现在城市都在走向智能和智慧，比如，智能电网、智慧交通、智慧医疗、智慧环保、智慧城市，这些都依托于大数据，可以说大数据是智慧的核心能源。截至 2020 年 4 月初，住房和城乡建设部公布的智慧城市试点数量已经达到 290 个；再加上相关部门所确定的智慧城市试点数量，我国智慧城市试点数量累计近 800 个，我国正成为全球最大的智慧城市建设实施国。充分运用互联网、大数据、人工智能等信息技术手段的智慧城市，面对疫情遭遇了严峻的考验，也在过程中凸显出其优势和短板。在这次新冠疫情暴发之初，浙江就通过大数据分析出，全省涉湖北的旅居人数超过了 30 万，预示着浙江将会有一定程度的疫情。智能交通系统能够对车辆进行快速检索与分析，可对重点车辆进行拦截报警，实现实时追踪和行驶轨迹预测。人员追踪系统可以追踪新冠肺炎确诊患者曾经的出行乘车记录、公共场所出入记录及接触人群等，锁定潜在病毒感染风险的人群，为防疫部门的追踪管理工作提供支持。湖北省应急物资供应链管理平台，针对抗击疫情急需的防护服、口罩、护目镜等物资的生产、库存、调拨、分配过程，进行了全程可视追踪、高效集中管控。

另外，作为国家的管理者，政府应该有勇气将手中的数据逐步开放，提供给更多有能力的机构组织或个人来分析并加以利用，以加速造福人类。比如，美国政府就筹建了一个 data.gov 网站，这是奥巴马任期内的一个重要举措：要求政府公开透明，而核心就是实现政府机构的数据公开。

企业的 CXO 们最关注的还是报表曲线的背后能有怎样的信息，他该做怎样的决策，其实这一切都需要通过数据来传递和支撑。在理想的世界中，大数据是巨大的杠杆，可以改变公司的影响力，带来竞争差异、节省金钱、增加利润、愉悦买家、奖赏忠诚用户、将潜在客户转化为客户、增加吸引力、打败竞争对手、开拓用户群并创造市场。

那么，哪些传统企业最需要大数据服务呢？抛砖引玉，先举几个例子：①对大量消费者提供产品或服务的企业（精准营销）；②做小而美模式的中长尾企业（服务转型）；③在互联网压力之下必须转型的传统企业（生死存亡）。

对于企业的大数据，还有一种预测：随着数据逐渐成为企业的一种资产，数据产业会向传统企业的供应链模式发展，最终形成“数据供应链”。这里尤其有两个显著的现象。①外部数据的重要性日益超过内部数据。在互联互通的互联网时代，单一企业的内部数据与整个互联网数据比较起来只是沧海一粟。②能提供包括数据供应、数据整合与加工、数据应用等多环节服务的公司会有明显的综合竞争优势。

对于提供大数据服务的企业来说，他们等待的是合作机会，就像微软总裁史密斯说的：“给我提供一些数据，我就能做一些改变。如果给我提供所有数据，我就能拯救世界。”

然而，一直做企业服务的巨头将优势不再，不得不眼看着新兴互联网企业加入战局，开启残酷竞争模式。为何会出现这种局面？从 IT 产业的发展来看，第一代 IT 巨头大多是 ToB 的，如 IBM、Microsoft、Oracle、SAP、HP 这类传统 IT 企业；第二代 IT 巨头大多是 ToC 的，如 Yahoo、Google、Amazon、Facebook 这类互联网企业。大数据到来前，这两类公司彼此之间基本是井水不犯河水；但是到了当前这个大数据时代，这两类公司已经展开了竞争。比如 Amazon 已经开始提供云模式的数据仓库服务，直接抢占 IBM、Oracle 的市场。这个现象出现的本质原因是：在互联网巨头的带动下，传统 IT 巨头的客户普遍开始从事电子商务业务，正是由于客户进入了互联网，所以传统 IT 巨头们不情愿地被拖入了互联网领域。如果他们不进入互联网，其业务必将萎缩。在进入互联网后，他们又必须将云技术、大数据等互联网最具有优势的技术通过封装打造成自己的产品再提供给企业。

以 IBM 为例，上一个 10 年，他们抛弃了 PC，成功转向了软件和服务，而这次将远离服务与咨询，更多地专注于因大数据分析软件而带来的全新业务增长点。IBM 执行总裁罗睿兰认为，“数据将成为一切行业当中决定胜负的根本因素，最终数据将成为人类至关重要的自然资源。”IBM 积极地推出了“大数据平台”架构，该平台的四大核心能力包括 Hadoop 系统、流计算（StreamComputing）、数据仓库（Data Warehouse）和信息整合与治理（Information Integration and Governance）。

另外一家亟待通过云和大数据战略而复苏的巨头公司 HP 也推出了自己的产品——HAVEn，一个可以自由扩展伸缩的大数据解决方案。这个解决方案由 HP Autonomy、HP Vertica、HP ArcSight 和惠普运营管理（HP OperationsManagement）四大技术组成，还支持 Hadoop 这样通用的技术。HAVEn 不是一个软件平台，而是一个生态环境。四大组成部分满足不同应用场景的需要：Autonomy 解决音视频识别的重要解决方案；Vertica 解决数据处理的速度和效率的方案；ArcSight 解决机器的记录信息处理，帮助企业获得更高安全级别的管理；运营管理解决的不仅仅是外部数据的处理，还包括了 IT 基础设施产生的数据。

个人的大数据这个概念很少有人提及，简单来说，就是与个人相关联的各种有价值数据信息被有效采集后，可由本人授权提供给第三方进行处理和使用，并获得第三方提供的数据服务。

举个例子来说明会更清晰一些。

未来，每个用户都可以在互联网上注册个人的数据中心，以存储个人的大数据信息。用户可确定哪些个人数据可被采集，并通过可穿戴设备或植入芯片等感知技术来采集捕获个人的大数据，比如，牙齿监控数据、心率数据、体温数据、视力数据、记忆能力数据、地理位置信息、社会关系数据、运动数据、饮食数据、购物数据，等等。用户可以将其中的牙齿监测数据授权给 XX 牙科诊所使用，由他们监控和分析这些数据，进而为用户制订有效的牙齿防治和维护计划；也可以将个人的运动数据授权提供给某运动健身机构，由他们监测自己的身体运动机能，并有针对性地制订和调整个人的运动计划；还可以将个人的消费数据授权给金融理财机构，由他们帮忙制订合理的理财计划并对收益进行预测。当然，其中有一部分个人数据是无须个人授权即可提供给国家相关部门进行实时监控的，如罪案预防监控中心可以实时地监控本地区每个人的情绪和心理状态，以预防自杀和犯罪的发生。

以个人为中心的大数据有这样一些特性。

数据仅留存在个人中心，其他第三方机构只被授权使用（数据有一定的使用期限），且必须接受用后即焚的监管；

采集个人数据应该明确分类，除了国家立法明确要求接受监控的数据外，其他类型数据都由用户自己决定是否被采集；

数据的使用只能由用户进行授权，数据中心可帮助监控个人数据的整个生命周期。

展望过于美好，也许实现个人数据中心将遥遥无期，也许这还不是解决个人数据隐私问题的最好方法，也许业界对大数据的无限渴求会阻止数据个人中心的实现，但是随着数据越来越多，在缺乏监管之后，必然会有一场激烈的博弈：到底是数据重要还是隐私重要？是以商业为中心还是以个人为中心？

2. 大数据的未来发展

未来大数据的身影应该无处不在，就算无法准确预测大数据终会将人类社会带向哪种最终形态，但相信只要发展的脚步还在继续，因大数据而产生的变革浪潮就会很快湮没地球的每一个角落。

比如，Amazon 的最终期望是：“最成功的书籍推荐应该只有一本书，就是用户要买的下一本书。”

Google 也希望当用户在搜索时，最好的体验是搜索结果只包含用户所需要的内容，而这并不需要用户给予 Google 太多的提示。

当物联网发展到一定规模时，借助条形码、二维码、RFID 等能够唯一标识产品，传感器、可穿戴设备、智能感知、视频采集、增强现实等技术可实现实时的信息采集和分析，这些数据能够支撑智慧城市、智慧交通、智慧能源、智慧医疗、智慧环保的理念需要，这些所谓的智慧将是大数据的数据采集来源和服务范围。

未来的大数据除了将更好地解决社会问题、商业营销问题、科学技术问题之外，还有一个可预见的趋势是以人为本的大数据方针。人才是地球的主宰，大部分的数据都与人类有关，要通过大数据解决人的问题。

比如，建立个人的数据中心，记录每个人的日常生活习惯、身体体征、社会网络、知识能力、爱好性情、疾病嗜好、情绪波动……换言之，就是记录人从出生那一刻起的每一分每一秒，将除了思维外的一切都存储下来，这些数据可以被充分利用：

医疗机构将实时监测用户的身体健康状况；

教育机构更有针对性地制订用户喜欢的教育培训计划；
服务行业为用户提供即时健康的符合用户生活习惯的食物和其他服务；
社交网络为用户提供合适的交友对象，并为志同道合的人群组织各种聚会活动；
政府能在用户的心理健康出现问题时实施有效的干预，防范自杀、刑事案件的发生；
金融机构能帮助用户进行有效的理财管理，为用户的资金提供更好的使用建议和规划；
道路交通、汽车租赁及运输行业可以为用户提供更合适的出行线路和路途服务安排；
……

当然，上面的一切看起来都很美好，但是否是以牺牲了用户的自由为前提呢？只能说新鲜事物在带来便利的同时也带来了弊端。比如，在手机普及之前，大家喜欢聚在一起聊天，自从手机普及后特别是有了互联网，人们不用聚在一起也可以随时随地沟通交流，这就是便利滋生的另外一种情形，人们慢慢习惯了和手机共度时光，人与人之间的情感交流仿佛永远隔着一张“网”。