跨境电商精英人才培养系列

跨境电子商务基础
跨境电子商务客服
跨境电子商务通关实务
跨境电子商务实务
跨境电子商务视觉设计
跨境电子商务物流管理
跨境电子商务数据化管理
跨境电子商务英语
跨境电子商务物流英语
商务英语函电

策划编辑：刘　建
责任编辑：高　宇
封面设计：华腾视觉 · 刘文东

ISBN 978-7-5635-6533-7

定价：38.00元

北京邮电大学出版社

跨境电子商务数据化管理

主编　王　慧

跨境电商精英人才培养系列

跨境电子商务数据化管理

KUAJING DIANZI SHANGWU SHUJUHUA GUANLI

主编　王　慧

北京邮电大学出版社
www. buptpress. com

智慧学习平台

跨境电商精英人才培养系列

跨境电子商务数据化管理

KUAJING DIANZI SHANGWU SHUJUHUA GUANLI

主编　王　慧

北京邮电大学出版社
www. buptpress. com

内 容 简 介

本书共八个项目，内容包括大数据认知、跨境电子商务数据分析、数据化选品与分析引流、跨境电子商务市场数据分析、跨境电子商务店铺经营数据分析、跨境电子商务营销数据分析、跨境电子商务数据报告、跨境电子商务物流数据分析。

本书可作为职业院校电子商务、跨境电子商务等专业的教材，也可供相关人员参考。

图书在版编目（CIP）数据

跨境电子商务数据化管理 / 王慧主编. -- 北京：
北京邮电大学出版社，2022.2

ISBN 978-7-5635-6533-7

Ⅰ.①跨… Ⅱ.①王… Ⅲ.①电子商务—运营管理—
职业教育—教材 Ⅳ.①F713.365.1

中国版本图书馆 CIP 数据核字（2021）第 217864 号

书　　名：跨境电子商务数据化管理
主　　编：王　慧
责任编辑：高　宇
出版发行：北京邮电大学出版社
社　　址：北京市海淀区西土城路 10 号（邮编：100876）
E-mail：publish@bupt.edu.cn
经　　销：各地新华书店
印　　刷：三河市龙大印装有限公司
开　　本：787 mm×1 092 mm　1/16
印　　张：10.5 插页 1
字　　数：217 千字
版　　次：2022 年 2 月第 1 版　2022 年 2 月第 1 次印刷

ISBN 978-7-5635-6533-7　　定　价：38.00 元

PREFACE 前言

跨境电子商务(简称“跨境电商”)是电子商务的一种特殊形式,通常指分属不同国家或地区的交易主体通过电子商务平台达成交易,进行支付结算,并通过跨境物流送达商品,最后完成交易的一种国际商业活动。跨境电商是电子商务应用的更高级形式,其充分利用现代信息技术,与实体经济加速融合,对人们的日常生产、生活和消费产生深刻影响。因为不同国家或地区的贸易双方均可通过互联网或者相关信息平台实现这一交易,所以跨境电商已日益成为在信息化、网络化、市场化、国际化条件下配置资源的重要途径,以及引领经济社会发展进步的一种重要力量。

数据分析与管理一直是电商行业从业人员必须要做的工作。从业人员通过对相关产品的不同数据或竞争对手产品数据的分析,可以得出相应的结论,从而为后续制定发展战略和规划提供依据。

对于跨境电商从业人员来说,掌握必备的跨境电商数据化管理知识与技能,对于工作的顺利开展有着十分重要的意义和作用。

为落实习近平新时代中国特色社会主义思想进课程进教材,积极培育和践行社会主义核心价值观,体现中华优秀传统文化等立德树人育人战略,推进教材、教法改革服务于人才培养,深化产教融合、校企合作,落实“双元”合作开发教材,及时将产业发展的新技术、新工艺、新规范纳入教材内容,围绕教学改革和“互联网+”职业教育发展等需求,编者编写了本书。

本书具有以下一些特点。

(1)内容新颖,案例丰富,紧扣“大众创业、万众创新”理念,关注跨境电商行业的发展动态,对大学生创新创业能力的培养起到引领作用。

(2)案例紧扣实际,基于我国跨境电商未来的发展重心将逐步向行业细分领域转移,所涉及的行业领域会越来越广,本书所选取的案例均为来自企业近几年一线的案例。

(3)有机融入课程思政内容,从国家政策、行业规范、从业素质等方面实现立德树人的目标。

(4)为拓展大学生认知面,本书特别设置了多个栏目,包括“案例导入”“小常识”“知识链接”等。同时,为了增强大学生的实践能力,每个项目最后还设置了项目实训。

本书推荐学时安排见下表。

项目序号	项目内容	学　时
1	大数据认知	2
2	跨境电子商务数据分析	4
3	数据化选品与分析引流	6
4	跨境电子商务市场数据分析	6
5	跨境电子商务店铺经营数据分析	6
6	跨境电子商务营销数据分析	4
7	跨境电子商务数据报告	2
8	跨境电子商务物流数据分析	4
总计		34

本书由浙江育英职业技术学院王慧任主编。

由于编者水平有限,书中难免存在疏漏和不足之处,敬请广大读者批评指正。

编　者

CONTENTS

目录

项目 1 大数据认知

知识目标

(1)了解大数据的概念。

(2)了解大数据的发展趋势。

(3)掌握大数据采集的基础设施。

(4)熟悉大数据分析工具与应用。

技能目标

(1)知道大数据采集的基础设施有哪些。

(2)会使用大数据分析工具进行简单的大数据分析。

重点及难点

重点：

(1)大数据的概念与特点。

(2)大数据采集的基础设施。

难点：

(1)大数据的发展趋势。

(2)大数据分析工具与应用。

【案例导入】

跨境电商在大数据时代如何挖掘自身长处

“2017 中国新外贸梦想节”在深圳罗湖体育馆举行。大会上精英云集，参会人员围绕新形势下的外贸格局为中国的新外贸提供发展新思路，寻求、开拓新方法。绍兴市柯桥区跨境电商行业协会会长江××等外贸精英更是受邀在大会上进行互动直播，以新外贸人的独特视角探讨新旧外贸的差异和进一步的发展规划。

在会议上，“跨境电商”成了几乎所有嘉宾都提到的热词。近几年来，随着传统出口贸易不断线上化、交易化，越来越多的外贸企业从网上获得商机。而作为推动经济一体化、贸易全球化的技术基础，跨境电商不仅冲破了国家间的障碍，使国际贸易走向无国界贸易，同时它也正在引起世界经济贸易的巨大变革。对企业来说，跨境电商构建的开放、多维、立体的多边经贸合作模式极大地拓宽了进入国际市场的路径，大大促进了多边资源的优化配置与企业间的互利共赢。

在外贸环境变化迅速的当下，“大数据”“云计算”等每一个词汇都在冲击着跨境电商的固有体制。由于全球化竞争加剧，前路迷茫更是跨境电商在新外贸形势下的典型表现。

为了探讨这一问题，让跨境电商能更好地在中国得到发展，江××等嘉宾纷纷表达了自己的观点，提出了新时代跨境电商的前进方向。

1. 大数据时代，跨境电商须挖掘自身优势

和旧外贸时代不同，新外贸环境中到处充满开放的元素。而以网络为媒介的跨境电商，更是在大数据时代的背景下展现出其透明的一面。在各种电商平台上，电商企业的各种信息能够直观地呈现给浏览者，包括其运营情况、征信情况等。这对于电商企业而言，既是机遇，也是挑战。这要求电商企业必须保证诚信经营，并且向自己的合作方呈现出一个切实可行的经营项目，否则就会因为数据的透明而暴露自己的缺陷，从而失去合作伙伴。但同时，信息的高度流通也使得精准合作成为可能，只要跨境电商企业准确地在平台上表现出自身产品的特色，就有可能获得需求者的青睐。大数据时代同样是一个“个性定制时代”，只有找准自己的优势，发掘自己的所长，才能使自身的收益最大化。

2. 新外贸时代，跨境电商应注重全球合作

这是一个全球化的时代，在这个大背景下，贸易的方向往往也是相互的。曾经，一家企业可以这样夸耀自己：“我们的产品销往世界各地。”但是如今，这样扁平化的商业结构明显已经不能适应当下的潮流。在新外贸时代，贸易结构是立体化的、多元化的。企业不再是单纯地向外输出产品，同样也在接纳来自世界各地的产品。这就要求跨境电商企业有更为广阔的视野和目标。电商企业不能只站在“地面上”规划蓝图，更需要“到天上去”，积极使用云计算等先进技术，“从天上往下看”，将目光放在更多的商品和企业上。在向外界输出产品的同时，也不能忽视了外国的优秀品牌和技术；在扩大海外市场的同时，也要汲取外界的优良技术，以提升自身的素质，在全球化的汹涌浪潮中获得立足之地。

3. 卖方市场已然到来，客户体验将成服务焦点

近十几年来，外贸市场环境的变化可谓天翻地覆。刚加入世界贸易组织（World Trade Organization，WTO）的中国曾享受到了不菲的人口红利，但是随着经济的发展，廉价的劳动力已经不足以支撑中国企业在外贸市场上驰骋。而大数据时代的到来，使得买方对商品的质量有了更高的要求。在市场竞争激烈的当下，如果企业只将目光集中在提升商品的质量上，那么无疑会使运营成本过高。因此，在卖方市场上，客户体验将成为跨境电商需要考量的一个指标。为客户提供优质的服务，能在一定程度上拉近合作双方的距离，尤其是通过网络进行沟通的合作双方，优质的服务能让对方感受到企业的诚意，从而促进合作顺利进行。

江××认为，现在这个时代很特殊，人们拥有全球化的环境、便捷的网络、化繁就简的跨境贸易程序，因此，这是一个跨境电商大有可为的时代。

“曾经是我们制定标准，我们挑选客户。那个时候全球经济好，你有货，客户就来抢。那个时候我们的劳动力还算低廉，所以在市场上也有优势。但是 2008 年金融危机后，情况就不一样了，原本的大客户都变成了小客户，客户碎片化的情况非常显著。”江××在直播会议上谈到，“因此，跨境电商需要思变。现在不再是我们挑选客户，而是客户来挑选我们，因此，我们需要读懂客户的需求，为客户提供贴身的服务，增强客户黏性。同时，我们还要善于利用工具更好地管理一些碎片化的信息。你不用，你的竞争对手在用，你就会落后。我们不能闭门造车，只有打开胸怀，选择最合适的发展方式，做好我们的强项，拉高我们的格局，才能在这个竞争激烈的市场争得一席之地。”

资料来源：https://www.xiaokeduo.com/news/3238.html，有改动。

1.1 大数据概述

1.1.1 大数据的概念与特征

1. 大数据的概念

诸多专家、机构从不同角度提出了对大数据的解释。当然，由于大数据本身具有较强的抽象性，目前国际上尚没有一个公认的定义。

维基百科认为，大数据是超过当前现有的数据库系统或数据库管理工具的处理能力，处理时间超过客户能容忍时间的大规模复杂数据集。

企业数据集成软件商 Informatica 认为，大数据包括海量数据和复杂的数据类型，其规模超过传统数据库系统进行管理和处理的能力。

亚马逊大数据科学家约翰·拉瑟(John Rauser)提到一个简单的定义：大数据就是任何超过了一台计算机处理能力的庞大数据量。

百度百科对大数据的定义为：无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合，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。

2. 大数据的特征

一般认为，大数据主要具有 5 个典型特征("5V"特征)，如图 1-1 所示。

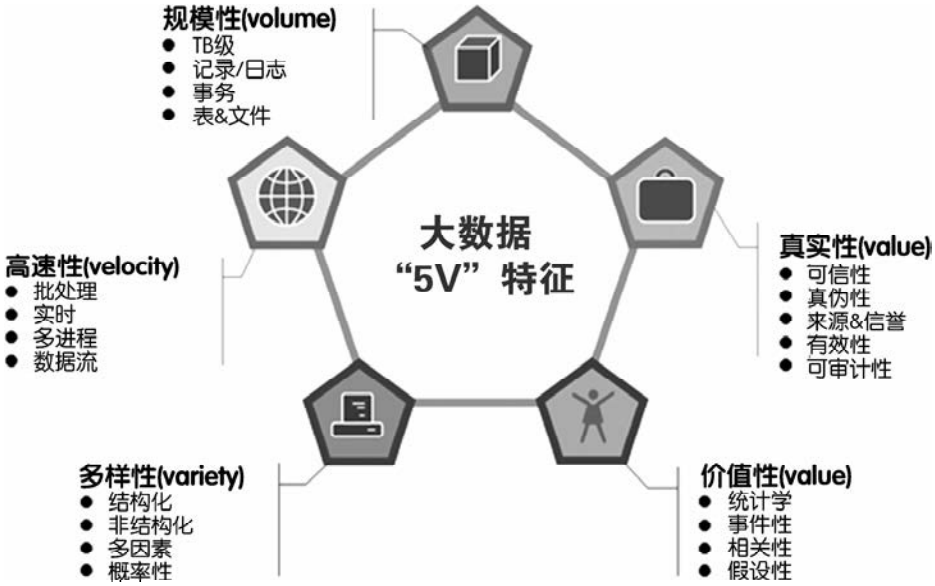

图 1-1　大数据的 5 个典型特征

(1)规模性。大数据的特征首先就体现为"数量大"，存储单位从过去的 GB 到 TB，直至 PB、EB。随着信息技术的高速发展，数据开始爆发性增长。社交网络(微信、微博、Facebook、Twitter)、移动网络、各种智能终端等都成为数据的来源。淘宝网近 4 亿会员每天产生的商品交易数据约为 20 TB；Facebook 约 10 亿用户每天产生的日志数据超过 300 TB。它们迫切需要智能的算法、强大的数据处理平台和新的数据处理技术来统计、分析、预测和实时处理如此大规模的数据。

那么，1 分钟到底会有多少数据产生呢？

①Google 收到超过 2 000 000 个搜索查询。

②Facebook 用户分享 684 478 条内容。

③Twitter 用户发送超过 100 000 条微博。

④苹果公司收到大约 47 000 个应用下载。

⑤Facebook 上的品牌和企业收到 34 722 个赞。

⑥Tumblr(汤博乐)博客用户发布 27 778 个新帖子。

⑦Instagram(照片墙)用户分享 36 000 张新照片。

⑧Flickr(雅虎网络相册)用户添加 3 125 张新照片。

⑨电子邮件用户发送 204 166 677 条信息。

⑩消费者在网购上花费 272 070 美元。

(2)高速性。与以往的档案、广播、报纸等传统数据载体不同,大数据的交换和传播是通过互联网、云计算等方式实现的,远比传统媒介的信息交换和传播速度快捷。大数据与海量数据的重要区别在于,大数据除了数据规模更大以外,对处理数据的响应速度也有更高的要求,实时分析而非批量分析,数据输入、处理与丢弃立刻见效,几乎无延迟。数据的增长速度和处理速度是大数据高速性的重要体现。

(3)多样性。广泛的数据来源决定了大数据形式的多样性。大数据大体可分为三类:一是结构化数据,如财务系统数据、信息管理系统数据、医疗系统数据等,其特点是数据间因果关系强;二是非结构化数据,如视频、图片、音频等,其特点是数据间没有因果关系;三是半结构化数据,如 HTML 文档、邮件、网页等,其特点是数据间的因果关系弱。

小常识

HTML

HTML 的英文全称为 hypertext markup language,即超文本标记语言,是一种标记语言。它包括一系列标签,通过这些标签可以将网络上的文档格式统一,使分散的互联网资源连接为一个逻辑整体。HTML 文本是由 HTML 命令组成的描述性文本,HTML 命令可以说明文字、图形、动画、声音、表格、链接等。

(4)价值性。价值性也是大数据的核心特征。现实世界所产生的数据中,有价值的数据所占比例很小。相比于传统的小数据,大数据最大的价值在于通过从大量不相关的各种类型的数据中挖掘出对未来趋势与模式预测分析有价值的数据,并通过机器学习方法、人工智能方法或大数据采集方法深度分析,发现新规律和新知识,从而运用于农业、金融、医疗等各个领域,最终达到改善社会治理、提高生产效率、推进科学研究的效果。

知识链接 大数据的核心价值

从业务角度出发,大数据的核心价值主要有以下 3 点。

①数据辅助决策:为企业提供基础的数据统计报表分析服务。分析师能够轻易获取

数据产出分析报告并指导产品生产和运营，产品经理能够通过统计数据完善产品功能和改善用户体验，运营人员可以通过数据发现运营问题并确定运营的策略和方向，管理层可以通过数据掌握企业运营状况，从而进行一些战略决策。

②数据驱动业务：通过数据产品、数据挖掘模型实现企业产品生产和运营的智能化，从而极大地提高企业的整体效能。最常见的应用领域有基于个性化推荐技术的精准营销服务、广告服务、基于模型算法的风控反欺诈服务、征信服务等。

③数据对外变现：通过对数据进行精心的包装，对外提供数据服务，从而获得现金收入。市面上比较常见的有各大数据公司利用自己掌握的大数据提供风控查询、验证、反欺诈服务，提供导客、导流、精准营销服务，提供数据开放平台服务，等等。

资料来源：http://www.itongji.cn/detail?type=1008，有改动。

(5)真实性。真实性其实就是数据的质量，海量数据并不一定都能反映用户真实的行为信息或者客观事物的真实信息。以网页访客数据为例，很多网站为了赚取更多的广告费用会使用机器人对广告进行点击，这样就产生了作弊流量，而这些流量并不能反映用户的真实需求。

今天的大数据已不只是“大”，真正有意义的是数据变得在线了，这远远比最初的“大”更能体现大数据的本质。从技术上看，大数据必然无法用单台计算机进行处理，必须采用分布式计算架构。分布式计算架构的特色在于对海量数据的挖掘，但它必须依托云计算的分布式处理、分布式数据库、云存储和虚拟化技术。从大数据的来源看，物联网、云计算、移动互联网、车联网、手机、平板电脑、个人计算机(PC)以及遍布地球各个角落的各种各样的传感器，无一不是数据来源或者承载的方式。大数据的核心价值在于对海量数据进行存储和分析。相比现有的其他技术而言，大数据的廉价、迅速、优化这三方面的综合成本是最优的。

小常识

车　联　网

车联网的概念源于物联网，即车辆物联网，是以行驶中的车辆为信息感知对象，借助新一代信息通信技术，实现车与车、人、路、服务平台等之间的网络连接，提升车辆整体的智能驾驶水平，为用户提供安全、舒适、智能、高效的驾驶感受与交通服务，同时提高交通运行效率，提升社会交通服务的智能化水平。

大数据的意义并不仅仅在于“大容量”，更重要的是通过对海量数据的整合、挖掘和分析可以创造出新的价值。所有跨境电商营销分析的第一步都是要找到目标受众，从而确定媒介投放策略。传统广告通过科学的手段探知受众，把握受众需求，以此做出市场预

判。互联网加剧了需求碎片化趋势，消费者的需求变得差异化、多元化、个性化，而同时互联网上的信息聚合和重构又提供了碎片重聚的可能。最为关键的是，大数据技术将这种可能变为现实，所有的这一切，都是为了把“消费者”还原成“整体的人”“丰富的人”，而不再是简单的“目标人群”。来源于社交网站的数据是大量的、鲜活的，反映了一个个具体网民的真实想法，体现了他们想做的事情。这些数据虽然价值密度低，但事关未来。通过分析、利用这些数据，企业营销人员能更加贴近消费者，深刻理解消费者的需求，还可以创造和引领消费者需求。

1.1.2 大数据的发展趋势与时代机遇

1. 大数据的发展趋势

自 2011 年起，大数据渐渐被人们了解，人们对大数据逐步形成相对完整的认知框架，并得出信息产业的发展具有三大趋势：数据成为资产，产业垂直整合，泛互联网化。这三大趋势的提出拓展了大数据主题的研究范围，为人们开辟了新的视角和逻辑来观察信息产业内企业的成长路径和投资价值。

(1)数据成为资产。在信息时代，数据将成为独立的生产要素。如果企业拥有某类相对完整、全面的数据，将可以实现“退可偏安一隅，进可跃马中原”。数据已经不是简单的“数字”了，它发生了很大的变化。

①成为工业化与信息化深度融合的关键枢纽。

②成为推动产业融合兼并的战略资产。

③成为各地方城市转换发展思路的新思维。

④成为推动企业跨行业转型的根据地。

⑤成为数学与工程实践结合的最佳演练场。

Google、亚马逊等互联网巨头积累了不同的数据资产。Google 为全世界的公开网页建立了庞大的索引系统；亚马逊网站上沉淀了大量的商品信息，成为互联网上最为庞大的商品数据库。不同的数据资产决定了不同的战略选择和商业模式。

拥有独一无二的数据资产的企业将会获得令人难以置信的发展速度，发展出令人叹为观止的商业模式。事实上，它们具备了颠覆、冲击其他行业的压倒性优势。

(2)产业垂直整合。新兴产业往往以垂直整合的态势开疆拓土，但产品成熟后，产业链上的分工专业化则会激发出惊人的创造力，并且成本也逐渐降低，优势逐渐转向水平分工格局。

信息产业中产业垂直整合趋势明显，是大数据效应改变产业竞争格局的一个缩影。在这个趋势下，越靠近终端用户的企业，在产业链中就拥有越大的发言权。微软和苹果公司的

此起彼伏是这个趋势最好的注脚。

过去人们购买计算机时，关注的是中央处理器（CPU）的主频、内存、操作系统等，而现在买 iPad 的人，多数是因为它比较“酷”，携带、使用比较方便，很少有人会问 iPad 的 CPU 是多大的。这标志着消费者的关注重点已经迁移到产品能否满足自身的个性化需求上。

在企业级市场也有相同的趋势，这个趋势的出现有两大原因：通用的平台型软件逐渐同质化，用户对自身业务的关注超过了对计算能力的追求。

（3）泛互联网化。互联网将逐步深刻影响社会的生产力和生产关系，泛互联网化可以理解为一切都在走向网络。现阶段，传统企业正在被互联网化，而下一阶段，人和物也将被互联网化。在泛互联网化范式中，强调终端、平台、应用“三位”加上“大数据”这“一体”。

大数据、终端、平台和应用都可以成为盈利的主要来源，在不同的发展阶段，盈利的主体也不尽相同。根据企业主要盈利来源的不同，可以把企业简单归类为图 1-2 所示的 5 种模式。

图 1-2　根据企业主要盈利来源不同划分企业模式

把数据当作资产，则更强调数据的战略意义。产业垂直整合趋势在数据运用层面就是通过收集大量的用户数据，更贴近用户，更理解用户，为其提供更适当的服务。泛互联网化驱动大数据飞轮效应的第一步是收集数据的重要渠道，利用没有泛互联网化的应用软件和硬件设备，企业将难以获得用户的行为数据。

2. 大数据的时代机遇

大数据价值被认知与被挖掘建立在一个前提下——数据化。不能将数据化等同于数字化，后者是指将模拟数据转换为二进制码，以方便计算机存储和分析，而前者则是指把日常生活、生产、商业等方方面面的现象转化为可制表分析的量化形式的过程。正是这个过程形成了各行各业的变革力量，因为这是大数据时代所独有的一种新型能力，以一种前所未有的方式，通过对海量数据进行分析，获得有巨大价值的产品和服务或深刻的洞见。中国工程院院士孙凝晖对此表示：“大数据在未来很可能会成为一个新的行业，而且大数据本身也超越了互联网行业，不仅仅是在网络上，生物基因本身也是大数据，各个物种的基因数据产生以后也会产生很多学术价值、商业价值。”这种说法并非没有依据。

从美国市场上已经发生的案例来看，互联网行业、商业智能与咨询服务领域、零售行业

受益最大，但医疗卫生、交通、物流甚至生物科技、天文等领域都开始“承认”大数据的价值。事实上，在美国各个行业和应用领域，大数据的应用已经遍地开花。互联网企业雅虎公司于2008 年年初便开始启用大数据技术，每天分析超过 200 PB 的数据，使服务变得更人性化，更贴近客户。雅虎将大数据技术应用于 IT 系统的方方面面，包括搜索、广告、客户体验和欺诈发现等。为了更深入地了解每一个客户，亚马逊不仅从每个客户的购买行为中获得信息，还将每个客户在其网站上的所有行为都记录下来，对这些数据的有效分析使得亚马逊对于客户的购买行为和喜好有了全方位的了解，对于其货品种类、库存、仓储、物流及广告业务都有着极大的效益回馈。


小常识

PB

PB 是计算机存储容量单位，1 PB＝1 024 TB＝2^{50} B。

大数据在医疗卫生领域的应用也正在爆发——通过大数据辅助癌症治疗，通过智能手机上的应用程序来监测病人的身体颤动，丹麦癌症协会甚至通过大数据来研究使用手机是否致癌等。

不得不提的还有零售行业。实际上，诸如沃尔玛、Tesco（英国零售商）等零售巨头已从大数据应用中获得了巨大的利益，也因此巩固了自己在业界长盛不衰的地位。

1.1.3 大数据时代的挑战

1. 大数据安全的隐患

虽然海量信息的集中存储会使数据的分析处理更加便捷，但在管理不当的情况下容易导致数据泄露、丢失或损坏，使企业利益遭受重大损失。

新媒体的发展影响着各类人群，其影响已从中国走向世界。越来越多的人生活在数据化、网络营销化、电商化、“微”化的世界中。在大数据时代，人们的一言一行几乎完全是透明的，个人隐私在网上被泄露成为亟须解决的问题。只要打开计算机，打开手机，人们在网上的一举一动都会被随时随地监视，监视结果信息会被汇报给企业，为企业描绘出一幅幅真实的且非常具有个性化的数据影像。

个人在使用网络进行各种社交、购物、保存私人物品时产生的数据，很容易被人盗用，导致自身的信息泄露；更害怕别有用心的人或者企业通过修改数据来进行欺骗，也就是数据造假。企业会害怕数据不公、数据垄断。个人隐私数据一旦被泄露（或被买卖），可能会对用户人身财产安全、企业安全甚至国家安全造成威胁。在隐私权和个性化之间永远需要一个平

衡，这个平衡一方面来自所谓的行业自律，另一方面来自消费者的自我保护，更重要的是来自政府第三方监管的跨界过程。

知识链接

大数据带来的好处

（1）生活方面。自助缴水电费、燃气费、电视费，汽车摇号、违章查询、公积金查询、手机代开发票、查询法院案子进展……这是运用大数据促进改善民生的典型事例。此外，大数据还被运用到智能家居中，如智能照明体系等。

（2）医疗方面。到目前为止，大数据最强大的应用之一就是电子医疗记录的收集。每一个患者都有自己的电子医疗记录，包括个人病史、家族病史、过敏情况以及所有医疗检测结果等。利用大数据收集患者信息，可以帮助人们尽早发现疾病，不但可以降低人们身体健康受损的风险，还可以帮助人们减少医疗支出。

大数据在医疗方面的另一个创新应用是可穿戴设备，这些设备能够实时分析汇报穿戴者的健康状况。可穿戴设备还能在医疗机构之外的场所使用，患者使用这些设备在家就能获知自己的健康状况，同时还能获得智能设备所提供的治疗建议。

（3）出行方面。人们的出行越来越离不开大数据的协助。运用电子地图，初来乍到的游客可以在生疏的城市自由行走；出租车司机经过语音导航，可以知晓前方路况，防止堵车或超速违章……

大数据仍是缓解交通压力的利器，它可以预测未来交通状况，为改善交通状况提供优化方案，这有助于交通部门把控交通，防止和缓解交通拥堵。

资料来源：https://www.sohu.com/a/346279373_100065429，有改动。

2. 信息共享动力不足影响企业效率

虽然各地电子政务云部署得如火如荼，但是实施方法、部署方式、安全保障等许多问题都还没有得到完善解决，电子政务在我国尚处于从概念清晰到务实行动的初级阶段。大量部门的数据如一个个信息孤岛，既给政府调度和公众办事带来了不便，又制约了数据活力的激发。

小常识

电 子 政 务

电子政务是指国家机关在政务活动中全面应用现代信息技术、网络技术以及办公自动化技术等进行办公、管理和为社会提供公共服务的一种全新的管理模式。广义的电子政务应包括所有国家机构，而狭义的电子政务主要包括直接承担管理国家公共事务、社会事务的各级行政机关。

3. 数据分析能力不够

大数据不仅仅意味着数据数量庞大，还代表着数据种类繁多、结构复杂，变化的速度也极快。随着大数据时代的到来，企业经营决策面临的最大挑战不再是缺少数据，而是数据太多。企业对大数据的应用，首先要对大数据进行处理才能实现。企业信息部门只有通过对大数据关系进行重新建构，赋予大数据新的意义，发掘有价值的数据，大数据才能为企业所利用，给企业营销管理提供决策支持，构筑企业核心竞争力。当前，大多数企业缺乏数据分析与处理的相关人才，导致企业无法在第一时间准确地获得与企业营销相关的信息，或已掌握相关信息，但无法有效分析和预测行业动态，最终导致企业在激烈的市场竞争中受挫。更根本的是，国内的企业长期以来对于大数据的价值没有充分的认识，也没有依赖大数据做决策的习惯，甚至很多企业忽视大数据的存在，所以，很多企业都没有长期的保留数据与应用数据的计划，这也使得大数据分析的前提难以满足。

4. 数据精准性与服务精准性不对称

尽管大数据确实能够发现、跟踪和分析消费者的每个显性变化，但无法全面把控消费者的内心活动。因为消费者的购买心理本来就是一个“暗箱”，其购买行为是由很多因素综合决定的，可能是心理，可能是价格，也可能是环境因素等。因此，尽管大数据能够提供精准的数字，但很难提供精准的预测。

综上所述，对于企业来说，大数据既是机遇也是挑战，大数据潜在的巨大价值必然会掀起一场商业模式和营销决策的深刻变革。在大数据时代，企业为了获得领先优势，必须转换思维，变革营销模式，充分、有效地利用大数据，挖掘其蕴含的附加价值，力求在瞬息万变的全球化经济环境中赢得竞争优势。

小常识

暗 箱 理 论

暗箱理论是指消费者的心理如同暗箱，人们只能看到消费者购买的外界条件（产品信息、价格信息和促销信息）和最终选择的结果。暗箱理论就是研究消费者行为的基本理论，即“5W1H”理论。

（1）what：购买什么。了解消费者想购买什么或购买了什么。

（2）who：谁参与购买行为。既要了解使用产品的是哪些人，又要弄清购买行动中的“购买角色”。

（3）when：何时购买。了解消费者发生购买行为的具体季节、时间甚至时点。

（4）where：在何地购买。了解消费者在哪里购买、在哪里使用。

（5）why：为何购买。了解和探索消费者行为的动机或影响其他行为的因素。

（6）how：怎样购买。了解消费者怎样购买、喜欢什么样的促销方式，还要搞清楚消费者对所购买的商品如何使用。

1.2 大数据采集

1.2.1 大数据采集的概念

大数据采集就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中提取潜在有用的信息和知识的过程。与大数据采集含义相近的词有数据融合、数据分析和决策支持等。大数据采集的定义包括以下几层含义。

(1)数据源必须是真实的、大量的、含噪声的。

(2)发现的是用户感兴趣的信息和知识。

(3)发现的信息和知识要可接受、可理解、可运用。

(4)并不要求发现放之四海皆准的信息和知识。

这里有必要对知识进行讲解。从广义上理解,数据、信息也是知识的表现形式,但是人们更愿意把概念、规则、模式、规律和约束等看作知识,把数据看作形成知识的源泉,好像从矿石中采矿或淘金一样。

原始数据可以是结构化的,如关系数据库中的数据;也可以是半结构化的,如文本、图形和图像数据;甚至是分布在网络上的异构型数据。发现知识的方法可以是数学的,也可以是非数学的;可以是演绎的,也可以是归纳的。发现的知识可以用于信息管理、查询优化、决策支持和过程控制等,还可以用于数据自身的维护。

因此,大数据采集是一门交叉学科,它把人们对数据的应用从低层次的简单查询提升到高层次的从数据中挖掘知识,提供决策支持。在这种需求的牵引下,汇聚了不同领域的研究者,尤其是数据库技术、人工智能技术、数理统计、可视化技术、并行计算等方面的学者和工程技术人员,一并投身到大数据采集这一新兴的研究领域,形成新的技术热点。

知识链接 大数据采集方法

1. 数据库采集

传统企业会使用传统的关系型数据库 MySQL 和 Oracle 等来存储数据。随着大数据时代的到来,Redis、MongoDB 和 HBase 等 NoSQL 数据库也常用于数据的采集。企业通过在采集端部署大量数据库,并在这些数据库之间进行负载均衡和分片来完成大数据采集工作。

2. 系统日志采集

系统日志采集主要是收集企业业务平台日常产生的大量日志数据，供离线和在线的大数据分析系统使用。

高可用性、高可靠性、可扩展性是日志采集系统所具有的基本特征。系统日志采集工具均采用分布式架构，能够满足每秒数百兆(MB)的日志数据采集和传输需求。

3. 网络数据采集

网络数据采集是指通过网络爬虫或网站公开 API 等方式从网站上获取数据信息的过程。

网络爬虫会从一个或若干初始网页的 URL 开始获得各个网页上的内容，并且在抓取网页的过程中不断从当前页面上抽取新的 URL 放入队列，直到满足设置的停止条件为止。这样可将非结构化数据、半结构化数据从网页中提取出来，存储在本地的存储系统中。

4. 感知设备数据采集

感知设备数据采集是指通过传感器、摄像头和其他智能终端自动采集信号、图片或录像来获取数据。

大数据智能感知系统需要实现对结构化、半结构化、非结构化的海量数据的智能化识别、定位、跟踪、接入、传输、信号转换、监控、初步处理和管理等。其关键技术包括针对大数据源的智能识别、感知、适配、传输、接入等。

资料来源：https://blog.csdn.net/dsdaasaaa/article/details/93661858，有改动。

1.2.2 大数据采集的基础设施

对于大多数企业而言，数据仓库过去一直是，未来也将仍然是企业级机构所不可或缺的关键性组成部分。一套针对可扩展性而精心设计的基础设施正是大数据能否真正发挥作用的关键所在。

1. 云数据中心

云数据中心是在云计算背景下，新的业务需求和资源利用模式与数据中心的完美结合。云模式已成为企业利用数据中心平台应对大数据挑战的重要方式。根据 IBM(国际商业机器公司)的数据报告，当前数据中心有 85%的运算能力存在闲置情况，50%～60%的数据中心 IT 负载可以采用云计算技术。可从以下 3 个方面讨论云计算对大数据采集的价值。

(1)提高效率。云计算帮助大数据平台降低复杂性，简化运维，提升资源活性和利用效率。云计算通过基于网络的服务交付，将硬件等基础架构融合为无形的 IT 资源，并借助负

载均衡、分布式计算等技术手段，实现 IT 服务的特色化交付。

(2)降低成本。云计算帮助云数据中心降低成本，有利于将更多资金投入增值业务中。由于采用了大量的虚拟化技术和统一的跨平台管理，云计算可以帮助运营商/企业用户节省大量的设施成本和软件许可费用。此外，云数据中心的资源利用率能够得到进一步提升，并且在负载均衡方面也有更出色的表现，从而最大化保护用户投资，实现产品或服务生命周期内价值最大化。

(3)满足个性化需求。云计算可支撑基于大数据的灵活高效的 IT 服务，满足多种个性化需求。借助云计算的分布式系统和虚拟化灵活调配资源，可以为大数据的各项分析、处理、挖掘提供高效灵活的 IT 服务支撑，满足用户个性化/定制化大数据挖掘、分析需求。

面对海量数据的增长，传统架构虽然能够进行扩充，但它面临着不能实现水平横向扩展的局限性，传统的 IT 架构和数据处理方式无法有效地应对大数据环境。数据的存储、计算、管理、分析等节点都需要适应大数据需求的方案，同时也要满足性能上的扩展。因此，基于数据中心的 IT 基础设施也必将从传统的数据中心向云数据中心转型。

小常识

云 计 算

云计算是分布式计算的一种，是指通过网络“云”将巨大的数据计算处理程序分解成无数个小程序，然后通过多部服务器组成的系统处理和分析这些小程序得到的结果并返回给用户。简单地说，云计算早期就是简单的分布式计算，解决任务分发问题，并进行计算结果的合并。因而，云计算又称为网格计算。通过这项技术，可以在很短的时间(几秒)内完成对数以万计的数据的处理，从而提供强大的网络服务。

2. 计算虚拟化

对大数据稍有了解的人会知道，Hadoop 是所有大数据解决方案中最具成长性的平台，它通过集群搭建起的高性能计算和存储平台，利用分布式架构对海量数据进行分析和处理。但并不是所有的企业都有足够的精力和能力去应对部署 Hadoop 带来的挑战，如涉及较为专业的计算机理论、数据分析理论，这给 Hadoop 的使用带来了诸多不便。

在这种情况下，引入虚拟化解决方案就成为破解这些难题的关键。服务器虚拟化，甚至基于计算、网络、存储各个模块的全面虚拟化，将有助于降低成本和提升集群系统的可用性和可靠性，避免 Hadoop 集群带来的昂贵成本负担，使得广大中型企业也可以实现大数据的分析和应用，而且可以提升大数据的服务价值。

此外，基础设施的全面虚拟化还可以顺应大数据几何级数增长的发展态势，从而从一开始就紧随业务和大数据价值挖掘的需求而不断前进，提升大数据价值。

3. 大数据存储

大数据与其他类型数据的区别主要是分析应用程序，大数据本身就意味着数量庞大的需要使用标准存储技术来处理的数据。由于这些数据缺乏一致性，让标准处理和存储技术无计可施，而且运营开销以及庞大的数据量导致无法用传统的服务器和存储区域网络（storage area network，SAN，是通过专用高速网将一个或多个网络存储设备和服务器连接起来的专用存储系统）方法来有效地处理。也就是说，大数据需要自己专用的平台，这同样是 Hadoop 可以派上用场的地方。

Hadoop 是一个开源分布式计算平台，它提供了一种建立平台的方法，这个平台由标准化硬件（服务器和内部服务器存储）组成，并形成集群，从而能够并行处理大数据请求。从存储方面来看，这个开源项目的关键组成部分是 Hadoop 分布式文件系统（Hadoop distributed file system，HDFS），该系统具有跨集群中多个成员存储非常大文件的能力。

HDFS 通过创建多个数据块副本，然后将其分布在整个集群内的计算机节点，这提供了方便可靠、极其快速的计算能力。从目前来看，为大数据建立足够大的存储平台，最简单的方法就是购买一套服务器，然后让 Hadoop 来完成余下的工作。

在传统的数据仓库中挖掘相似数据集，一般都在一个单独的存储设备上进行，但这种方法对处理能力和存储容量的可扩展性来说已经不是最优的选择了。下面介绍几种适合存储大数据的方法。

（1）横向扩展网络附属存储（network attached storage，NAS）。横向扩展 NAS 是文件级别的访问存储器，它由多个连接在一起的存储节点构成，而且其存储容量和处理能力会随着节点的增加而提升。同时，它支持数十亿文件和 PB 级存储容量的并行文件系统，允许把不同位置的大量数据连接起来。

横向扩展 NAS 产品主要包括以下几个。

①EMC Isilon 及其 OneFS 分布式文件系统。

②HDS 的 Cloudera Distribution Hadoop（CDH）Cluster 基准体系架构。

③Data Direct Networks hScaler Hadoop NAS 平台。

④IBM 的 SONAS（Scale out Network Attached Storage）。

⑤HP 的 X9000。

（2）对象存储。对象存储有可能替代传统的树形文件系统，是适合存储大数据的。对象存储支持平行的数据结构，所有文件都有唯一的 ID 标识，类似于网上的域名系统（domain name system，DNS）。在平行的文件系统结构中处理大量的对象比在垂直的文件系统结构中要简单得多。

对象存储产品越来越多地支持大数据分析环境，其产品主要有 Scality 的 RING 体系结构、Dell 的 DX、EMC 的 Atmos 平台等。

（3）Hyperscale。Hyperscale 计算机/存储体系结构被多家大型公司使用，包括

Facebook、Google 等。Hyperscale 使用许多相对简单、常见的基于硬件的直连式存储计算机节点，来提高大数据分析环境的性能，如 Hadoop。

与传统的企业级计算和存储架构不同，Hyperscale 在完整的计算机/DAS（开放系统的直连式存储）节点上进行冗余备份。如果一部分节点遇到故障，失败的任务将会交给另一个备份节点，整个出故障的单元都将被替换。

4. 网络虚拟化

在虚拟数据中心，物理设备不再是数据中心的重点。虚拟化的运用使得物理功能被抽离出来，相关软件和数据运行在虚拟设备上，资源将基于策略得到优化调配。此外，特定工作任务也能脱离物理局限得以处理。同时，大量的、高速增长的、多维的结构化数据及非结构化数据将能更加动态、灵活地被管理、分析。据统计，目前虚拟服务器和实体服务器数量的比例为 11∶1，而这一比例正随着用户对虚拟数据中心需求的增长而增长。

尽管在虚拟化市场上，以 VMware 为代表的国外厂商起步早、投入大、市场规模大，但国产虚拟化厂商显然更能读懂国内用户的心。国产虚拟化厂商将更好地建设虚拟化数据中心，从而推动大数据能量倍增。网络虚拟化的特点如下。

（1）服务器虚拟化双机集群（high available，HA）高可用性保障由虚拟化软件直接提供，不需要额外的昂贵的专用备份软件。

（2）不需要主服务器与备份服务器的 CPU、内存、磁盘等硬件配置保持一致性，也就是说，可以在主服务器采用英特尔（Intel）处理器，而备份服务器采用 AMD 处理器的情况下，实现 HA 高可用性保障。

（3）主服务器和备份服务器可同时运行不同业务，不需要备份服务器的资源闲置，只要备份服务器上的冗余资源足以满足主服务器上的业务系统的最低需求，就可以实现 HA 高可用性保障。

（4）支持虚拟机 CPU 资源控制。能够实现灵活的资源配置，还能支持虚拟机批量操作，可大幅提升管理效率。

在容错方面，如果主服务器出现非计划宕机，可以让备份服务器自动检测到宕机故障，并自动将业务系统迁移到备份服务器上继续运行，从而保障数据的安全性和完整性，保证业务的连续性和一致性，避免业务终端损坏而带来的灾难性后果。

小常识

宕　机

宕机是指操作系统无法从一个严重系统错误中恢复过来，或系统硬件层面出现问题，以致系统长时间无响应，而不得不重新启动计算机的现象。它属于计算机运作过程中的一种正常现象，任何计算机都会出现这种情况。

1.3 大数据分析工具与应用

1.3.1 分布式平台

1. 分布式平台概述

分布式平台即分布式系统，是建立在网络上的系统软件，更是处理大数据的根基，而 Hadoop 就是著名的分布式平台。

云计算包含两个部分，即分布式文件系统（如 Google 的文件系统 GFS）和分布式表格系统（如 Google 的 Bigtable）。其中，分布式文件系统实现可靠、高效的数据存储和处理；分布式表格系统在分布式文件系统的基础上实现表格的各种处理逻辑，如查询、修改、扫描等。

此外，鉴于开发和调试分布式程序有比较大的难度，实现高效的分布式程序挑战更大，因而云计算还有一个分布式计算系统 MapReduce（一种编程模型）。通过它，云计算上的分布式程序开发就会变得容易很多，运行效率也大大提升。MapReduce 既可以运行在分布式表格系统上，也可以直接运行在分布式文件系统上，达到很高的并行度，获得很好的效率。

（1）单一主控机＋多工作机。云计算系统一般采用的是单一主控机＋多工作机模式，工作机实现数据的存储、读写、分析、处理等，主控机保存部分或全部元数据，实现工作机的任务分配、状态监控、负载平衡、故障监测和故障恢复等。

主控机常常使用类似机制监控工作机的状态，向工作机定期发放“任务”，工作机在“任务”有效期（如几十秒）内才进行工作，“任务”失效后则停止工作。如果主控机发现某个工作机在过去一段时间内没有响应或者出现其他异常，则不再向该工作机发放新的“任务”，并在旧的“任务”到期后重新分配该工作机上的任务。这使得主控机可以发现有故障的工作机并将其从系统中剔除，并在适当的时候采取措施以避免数据丢失或者任务失败等。

（2）单一主控机＋几个辅主控机。如果没有其他措施，则云计算系统的单一主控机会成为整个系统的单点。为了避免出现这种现象，云计算系统通常还有一个分布式选举系统（如 Google 的 Chubby），主控机也不再是单一主控机，而是单一主控机和几个辅主控机，辅主控机保持着对主主控机的准同步，一旦主主控机出现故障，则其中一个辅主控机就会被“选举”并升级成为主主控机。这种“选举”和升级通常需要若干秒的时间，但由于工作机在“任务”有效期内即使没有主控机也会继续工作，且应用程序对主控机的访问是通过名字而不是 IP 地址，因此上层应用程序通常看不到这种切换，或者只是一个短暂的停顿。

2. 分布式文件系统

云计算的分布式文件系统是整个云计算的基石，它用来提供上层表格系统所需的可靠和高效的数据存储。它有以下几条假设。

(1)容错与故障自动恢复是“基因”。由于整个分布式文件系统由许多廉价计算机组成，机器发生故障是常事而非例外，因此系统需要不停地进行自我检测和监控，以发现故障机器并自动恢复。

(2)系统存储大尺寸文件。整个分布式文件系统存储着数百万甚至数千万的 100 MB 或更大尺寸的文件，而不是数十亿的 KB 尺寸小文件。分布式文件系统虽然也支持小文件的创建、读写，但效率不高。

(3)文件的主要修改是追加。分布式文件系统支持高效的大尺寸数据追加，特别是来自多个用户的、无锁并发追加，虽然也支持小尺寸的数据追加和改写，但效率不高。

(4)高效的大尺寸数据顺序读。大尺寸数据的顺序读相对高效，小尺寸数据的随机读相对比较低效。

(5)持续可用的网络带宽比低的单次读写延时更加重要。多数上层应用程序对数据吞吐量有较高的要求，但对单次读写时间没有很高的要求。保持持续可用的网络带宽比保证每次读写的低延时有更大的意义。

在云计算的分布式文件系统中，数据被分成固定大小的块。由于可靠性和性能的需求，每个数据块在系统中有若干份副本，并保存在不同的工作机上。此外，这些副本所在的工作机通常位于不同的机架和不同的网络交换机中，因此，一个机架或交换机故障不会导致数据不可用。

把多个副本分布到不同交换机上，可以进一步提高数据读出的可用网络带宽，增加数据读出的性能。但这样做增加了写入时在不同交换机之间传输的数据量，提高了写入成本。由于对数据的读取远远多于对数据的写入，这种做法提高了系统的总体性能。

与云计算架构的其他子系统一样，云计算的分布式文件系统采用“单一主控机＋多工作机”的结构。其中，工作机保存数据块的副本，主控机保存文件和目录的名字空间、文件到块的映射、当前工作机列表、块副本在当前工作机上的分布等。此外，主控机还记录了工作机的数据块大小、可用磁盘空间，数据读写次数等，并在必要的时候进行块迁移，以便实现负载的相对平衡。

云计算的分布式文件系统还提供了客户端库，应用程序通过客户端库访问文件数据。例如，当客户端需要读出一个文件从某个位置开始的数据时，客户端库通过询问主控机获得该文件的指定位置所在的块以及该块所在的工作机列表，再向其中的一个工作机发起读取块的请求，工作机读出指定的数据后返回给客户端库，之后客户端库再返回给应用程序。

3. 分布式计算系统

众所周知，云计算属于分布式系统。在云计算分布式系统中，网络延时(毫秒级)远远大

于单机系统内延时(微秒级),加上部件的不可靠性以及节点之间较松的耦合度和异构性,导致高效并行程序的设计和实现难度极大,阻碍了普通程序员使用云计算系统。

为了解决这个问题,Google 把 MapReduce 模型成功地应用到云计算系统中,极大地降低了云计算系统应用程序的开发难度,并且提高了云计算系统的并行度和运行效率,这就是云计算的分布式计算系统,其优点如下。

(1)可容错。由于整个作业被切分成许多任务,当工作机发生故障时,再次执行对应的 Map(映射)和 Reduce(归约)任务即可。主控机则定期记录检查点,一旦主控机异常,新的主控机读入最后一次检查点,则整个作业可以在最后一次检查点的基础上继续执行。

(2)效率高。云计算分布式计算系统的主控机根据用户设置自动把作业切分为许多任务,然后以按需分配的方式分配任务到所有的工作机上,每个工作机完成一个任务后就报告给主控机,主控机给该工作机再分配一个任务,如此反复。

(3)适于异构机群。按需分配任务的方式使得每个工作机的计算能力都能得到最大限度的发挥,计算能力高的工作机执行更多的任务,计算能力低的工作机执行较少的任务。

(4)应用程序开发者不需要设计、编写和调试并行程序。开发者只需要设计、编写和调试普通的串行程序,调试通过后提交到云计算系统,由云计算分布式系统框架把它们分发到成百上千台计算机(云计算系统的工作机)上运行,并汇总和返回运行后的结果。

1.3.2 关键技术 Hadoop 云计算

只有掌握了更高效的数据处理方式,企业才能在跨境电商营销中占得先机。因此,了解 Hadoop 技术是企业运用大数据的基础与核心。该技术已被广泛用于 Google、Yahoo!(雅虎)、Facebook、eBay 等网络服务中。

1. Hadoop 的基本概念

Hadoop 可以说是最常见的分布式计算,是一个开发和运行处理大规模数据的软件平台,是 Apache 的一个用 Java 语言实现开源软件框架,实现在大量计算机组成的集群中对海量数据进行分布式计算。

Hadoop 框架中最核心的设计就是 HDFS 和 MapReduce。HDFS 提供了海量数据的存储,MapReduce 提供了对数据的计算。

Hadoop 的集群主要由 NameNode、DataNode、Secondary NameNode、JobTracker 以及 TaskTracker 组成,其作用如下。

(1)NameNode。NameNode 是 HDFS 的守护程序,负责记录文件是如何被分割成数据块的,以及这些数据块分别被存储在哪些数据节点上。它的主要功能是对内存和 I/O 端口(输入/输出端口)进行集中管理。

一般来说，NameNode 所在的服务器不存储任何用户信息，不执行计算任务，以避免这些程序降低服务器的性能。如果从服务器宕机，Hadoop 集群仍旧可以继续运转。但由于 NameNode 是 Hadoop 集群中的一个单点，如果 NameNode 服务器宕机，那么整个系统将无法运行。

(2)DataNode。集群中的每个从服务器都运行一个 DataNode 后台程序，这个后台程序负责把 HDFS 数据块读/写到本地的文件系统中。当需要通过客户端读/写某个数据时，先由 NameNode 告诉客户端去哪个 DataNode 进行具体的读/写操作，然后客户端直接与这个 DataNode 服务器上的后台程序进行通信，并对相关的数据块进行读/写操作。

(3)Secondary NameNode。Secondary NameNode 是一个用来监控 HDFS 状态的辅助后台程序，就像 NameNode 一样，每个集群都有一个 Secondary NameNode，并且部署在一台单独的服务器上。

Secondary NameNode 不同于 NameNode，它不接收或记录任何实时的数据变化，但是它会与 NameNode 进行通信，以便定期保存 HDFS 元数据的快照。由于 NameNode 是单点的，Secondary NameNode 的快照功能可以将 NameNode 的宕机时间和数据损失降到最低。同时，如果 NameNode 发生问题，Secondary NameNode 可以及时地作为备用 NameNode 使用。

(4)JobTracker。JobTracker 后台程序用来连接应用程序与 Hadoop。用户代码提交到集群以后，由 JobTracker 决定哪个文件将被处理，并且为不同的任务分配节点。同时，它还监控所有运行的任务，一旦某个任务失败了，JobTracker 就会自动重新开启这个任务，在大多数情况下，这个任务会被放在不同的节点上。每个 Hadoop 集群只有一个 JobTracker，一般运行在集群的主控机上。

(5)TaskTracker。TaskTracker 与负责存储数据的 DataNode 相结合。其处理结构也遵循主/从架构，JobTracker 位于主节点，统领 MapReduce 工作；而 TaskTracker 位于从节点，独立管理各自的任务。每个 TaskTracker 负责独立执行具体的任务，而 JobTracker 负责分配任务。

虽然每个从节点仅有唯一的 TaskTracker，但是每个 TaskTracker 可以产生多个 Java 虚拟机(Java virtual machine，JVM)，用于并行处理多个任务。TaskTracker 的一个重要职责就是与 JobTracker 交互，如果 JobTracker 无法准时获取 TaskTracker 提交的信息，JobTracker 就判定 TaskTracker 已经崩溃，并将这个任务分配给其他节点处理。

小常识

Java 虚拟机

Java 虚拟机是运行所有 Java 程序的抽象计算机，是 Java 语言的运行环境，它是 Java 最具吸引力的特性之一。虚拟机是一种抽象化的计算机，通过在实际的计算机上仿真模拟各种计算机功能来实现。Java 虚拟机有自己完善的硬件架构，如处理器、堆栈、寄存器等，还具有相应的指令系统。Java 虚拟机屏蔽了与具体操作系统平台相关的信息，使得 Java 程序只需生成在 Java 虚拟机上运行的目标代码（字节码），就可以在多种平台上不加修改地运行。

2. Hadoop 的发展现状

由于 Hadoop 优势突出，基于 Hadoop 的应用已经遍地开花，尤其是在互联网领域。Facebook 借助集群运行 Hadoop，以支持其数据分析和机器学习；百度则使用 Hadoop 进行搜索日志的分析和网页数据的挖掘工作；淘宝的 Hadoop 系统用于存储并处理电子商务交易的相关数据；中国移动通信有限公司研究院基于 Hadoop 的“大云”系统用于对数据进行分析并对外提供服务。

随着互联网的发展，新的业务模式还将不断涌现，Hadoop 的应用也会从互联网领域向电信、电子商务、银行、生物制药等领域拓展。相信在未来，Hadoop 将会在更多的领域中扮演“幕后英雄”，为人们提供更加快捷优质的服务。

总而言之，Hadoop 是一个可以在普通通用硬件集群上进行大规模数据处理的优秀工具。但是如果希望处理动态数据集、点对点分析或者图形化数据结构，那么 Google 已经展示了大大优于 MapReduce 范型的技术选择。毫无疑问，Percolator、Dremel 和 Pregel 将成为大数据的新“三巨头”，正如 Google 的老“三巨头”：GFS、GMR 和 BigTable。

1.3.3 云存储的重要性

随着物联网社交化、自带设备（bring your own device，BYOD）等技术的广泛应用，数据呈现爆炸性增长。这不仅对存储的性能及容量提出了苛刻的考验，还要求存储具备快速的数据检索与分析能力，以即时获取关键价值信息。在未来，云存储将是大数据和分析领域最大的基础设施分支之一。

小常识

BYOD

BYOD 指携带自己的设备办公，这些设备包括 PC、手机、平板电脑等（而更多情况下指手机或平板电脑这样的移动智能终端设备）。可以在机场、酒店、咖啡厅等随时随地登录公司邮箱、在线办公系统，不受时间、地点、设备、人员、网络环境的限制，BYOD 向人们展现了一个美好的未来办公场景。

存储和网络是 Hadoop 集群性能的重要保证。在 Hadoop 集群中，万兆位以太网带来的带宽增长是导入和复制（在多台服务器之间）大型数据集的关键，英特尔的融合网络适配器提供了高吞吐量连接，同时，英特尔 SATA 固态硬盘为原始存储提供了高性能、高吞吐量存储选择。

为提高效率，存储往往需要支持其他高级能力，如压缩、加密、自动数据分层、重复数据删除、纠删码和自动精简配置等，现有的英特尔至强处理器已经支持这些功能。随着大量 IT 厂商的加入，Hadoop 的商用版本呈现增长趋势，众多厂商都推出了自己的 Hadoop 版本，并集合了其他 Hadoop 项目的基本堆栈，可与数据仓库、数据库和其他数据管理产品集成。

英特尔 Hadoop 发行版免费版 V2.2 为最终用户和应用提供商提供了一个功能强大、方便易用的大数据入门平台，而且免费版和企业版共用相同的核心代码，免费版也包含所有核心增强功能，不过免费版在节点数和系统存储容量上有所限制。英特尔 Hadoop 发行版免费版具有以下 4 个特点：在稳定性和易用性方面做了优化；对英特尔的平台做了特殊的优化，这个软件包在性能上和效率上是有优势的；在算法和结构上做了调整，也就是对即时性的优化，使其能够做到即时的数据处理；与中国的用户合作，对行业应用做了特殊的调整和优化。

项目实训　了解大数据

【实训目的】

（1）掌握大数据的基础知识。

（2）掌握独立学习及独立思考的能力。

【实训内容】

（1）按每组 3～5 人进行分组，每个小组设组长一名，负责统筹工作。每个小组在组长的

带领下完成对大数据相关知识的复习。

(2)以小组为单位,查找关于大数据的课外资料。

【实训步骤】

(1)以小组为单位,组长带领组员完成大数据相关知识的复习。

(2)以小组为单位,组员查找关于大数据的课外资料并形成书面文件。

思考与练习

1. 大数据的发展趋势是怎样的?
2. 大数据时代的挑战是什么?
3. 大数据采集基础设施是什么?
4. 分布式计算的优点有哪些?