摘要：本文先梳理大数据产生的背景、概念与范畴、技术概貌，然后列举国内外知名的互联网公司对大数据发展的战略部署与实际应用，比较国内企业与国外企业在大数据技术应用方面的异同与差距，在此基础上对大数据业务的未来发展作出分析展望。

# 国内外主要互联网公司大数据布局与应用比较研究

文｜官建文　刘振兴　刘扬

## 大数据的由来

2001年，高德纳（Gartner）公司的一份研究报告首次出现“大数据（Big Data）”概念的提法。时至今日，虽然对“大数据”一词的定义说法不一，但越来越多的研究机构和网络媒体开始关注它。大数据正成为继云计算（Cloud Computing）之后新的热词。同云计算一样，大数据虽然也看不见摸不到，却与今日的信息技术发展如影随形，并潜行于当前的信息生产、加工、交换过程之中，我们已经享受到的某些信息服务，如在社交网站看到的是自己想关注甚至是感兴趣的广告而看不到不想关注的广告，这其实是大数据技术的功劳。

### 1. 发展背景

后Web2.0时代，移动互联网的带宽不断提高与智能设备销售量不断上升，互联网业迎来了“云计算”和“大数据”。世界经济论坛一份有关大数据的研究报告称，每天全球几十亿人使用计算机、GPS设备、电话和医疗设备，产生海量的数据信息。这些用户大部分来自发展中国家，他们的需求和习惯尚未被真正理解，如果能够借助大数据相关技术分析和挖掘数据背后的信息，将有助于认识需求、提供预测和防范危机。另有评述说，美国的汽车保有量是中国三倍，而其车祸死亡人数仅为中国的一半，这得益于信息社会的数据革命之功。

毫无疑问，现在我们比历史上任何时候拥有的数据信息都要多。这些数据来源各式各样：收集气候变化的传感器，社交媒体上的消息，数字照片和视频，交易记录，移动电话的GPS信号等等。中国移动研究院在一份简报中称，随着全球信息化的进程加快，数据量的增加已经到达了前所未有的速度，2011年创造的信息数据达到180亿GB，而且每年以60%增加，到2020年全球一年产生的数字信息将达到35ZB，相当于350万亿GB。数据在持续地增多变大，多到现有数据技术无法分析处理，我们需要专门来解读这些海量数据的技术，这就是“大数据技术”。

### 2. 基本概念

如同高德纳公司的报告里提到的那样，业界普遍认同所谓“大数据”具有明显的“3V特征”：量级（Volume），速度（Velocity）和多样性（Variety）。大数据普遍具有量级大，要求处理速度快，数据本身具有丰富的多样性。在甲骨文公司和中国移动研究院的相关研究文档里，都追加了第四个V——Value，价值，而IBM在其相关文档中给出的第四个“V”则是真实性（Veracity）。

基于此，大数据可以被定义为：以新数据处理技术为手段，在海量、结构复杂、内容多样的数据集中，以较快速度解析出规律性或根本性的判断、趋势或预见。更为简单地说，是数据集太大以至于传统数据库软件无法处理，所以称为“大数据技术”。

从数据生成类型上区分，大数据可分为交易数据、交互数据和传感数据；从数据来源上分，大数据可分为社交媒体、银行/购物网站、移动电话和平板电脑、各种传感器/物联网等等；从数据格式可以分为文本日志、整型数据、图片、声音、视频等；还可从数据关系上区分为结构化数据（如交易流水帐）和非结构化数据（如图、表，地图等）；从数据所有者可分为公司尤其巨型公司数据、

政府数据、社会数据——网络数据。

根据美国白宫的“大数据开发计划”中的说，大数据开发也可指“从庞大而复杂的数字数据中发掘知识及现象后的本质（extract knowledge and insights from large and complex collections of digital data）”。同时也看到，现在所讨论的大数据并不仅仅是数据尺寸的变大，它还可以被视作一个机会，籍此可以在新的正在生成的数据和内容中找到本质的东西，从而使商业运作更敏捷，帮助回答一些此前无法预知的问题。

### 3.主要技术概要

大数据的提出是为了解决现有数据技术无法满足快速增多、日益复杂化的数据集合，因此基于大数据的技术涉及层面较广，至少包括如下一些现有技术的综合运用。关联规则学习、分类、分组分析、众包技术、数据异构与同构、机器学习、自然语言处理、神经网络、模式识别、预测模型、情态分析、信号处理、时序分析和可视化处理等。

上述每一项技术如果展开来说，需要写很多篇文章来讨论。如关联规则学习，是数据挖掘的一个重要课题，用于从大量数据中挖掘出有价值的数据项之间的相关关系，由此产生了对基于大数据的推荐系统的应用研究。再如机器学习，机器学习算法是从数据中自动分析获得规律，并利用规律对未知数据进行预测的算法。因为学习算法中涉及大量的统计学理论，机器学习与统计推断学联系尤为密切，也被称为统计学习理论。算法设计方面，机器学习理论关注可以实现的、行之有效的学习算法。很多推论问题属于无程序可循难度，所以部分的机器学习研究是开发容易处理的近似算法。

大数据的具体化、实例化的应用离不开Apache Hadoop项目，一种开源、可扩展、分布式的应用计算架构。它包括Common、Distributed File System、MapReduce三个组件部分。Hadoop的Map/Reduce框架是一种主/从架构，机群中有单一的主服务器以及若干个从服务器，在每个节点都有一个从服务器，这些分布式的节点协同工作，共同完成一个整体的大数据处理任务。

## 国外主要互联网公司大数据战略布局与应用

大数据技术与业务发展，仍然以欧美国家大型IT公司为主进行。像上一拨“云计算”的热潮一样，大数据日渐成为IT厂商竞相抢占的制高点，图1是2011年大数据厂商的收益分析，数据来自Wikibon。

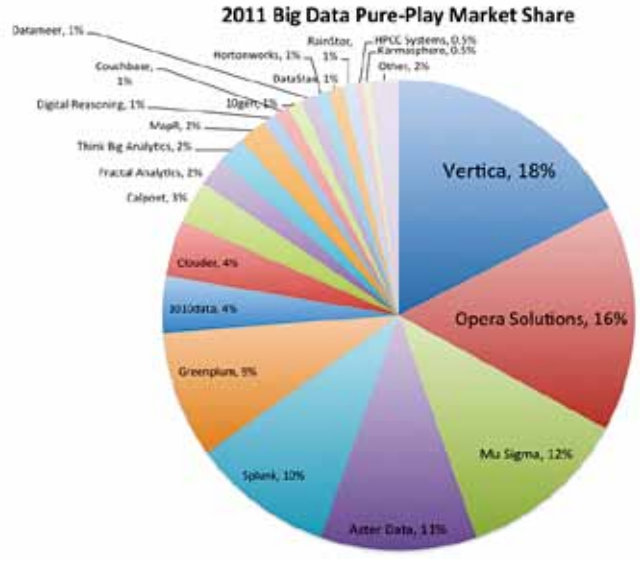


图1 大数据厂商2011年市场份额

### 1.国际商用机器（IBM）

IBM是商业分析和大数据技术的最活跃厂商之一。早在大数据概念进入媒体视野之前，IBM就提出“智慧地球”的说法，其核心是把“智慧”嵌入系统和流程之中，使服务的交付、产品开发、制造、采购和销售得以实现，使亿万人生活和工作的方式变得更加智慧。现在，大数据技术为IBM提供了一种实现途径。近年来，IBM先后投资了SPSS、Clarity、OpenPages、i2、Algorithmics等公司用以开发其商业分析解决方案，为客户提供预知判决、防范诈骗、风险和威胁的能力。此外，IBM雇佣了近9000名具有专业行业知识的资深分析咨询师，建立起了由8个全球分析解决方案中心联接起来的网络。

IBM大数据平台建立在开源的Apache Hadoop之上。通过向用户提供分析的整合手段从而理解信息以求得更好的商业效益，此平台能够使数据密集型应用软件更方便地管理和分析PB级大数据。IBM正在扩展其大数据平台以使其能在Hadoop的其它运营平台上运行，首先将推广至Cloudera。Cloudera对于Hdaoop社区的发展作出了重大贡献，同时也较早地为金融服务、政府、通信、媒体、零售、能源、医保等行业的客户提供了基于Hadoop的系统。Hadoop的Cloudera用户现在能够使用IBM大数

据平台进行复杂的数据分析，建立新一代的软件应用程序。

### 2.甲骨文（Oracle）

甲骨文公司在官方文档中将自身描述为“第一个为企业提供完整、集成的大数据全面解决方案的厂商”。它将大数据来源划分成为三类：1）传统企业数据，如CRM系统，ERP系统，在线交易数据等；2）机器生成/传感器数据，如呼叫记录，网络日志，智能度表，设备日志等；3）社交数据，如用户反馈系统，微博和校友录等等。甲骨文认为大数据对企业来说非常重要，可以帮助企业更深刻和透彻地理解商业行为，进而为改进服务，提高竞争力和更好地创新提供帮助。

甲骨文将大数据平台的行为概括为：数据获取、数据组织和数据分析。并为这三个阶段开发了不同的产品，而这些产品又与其推出的“大数据机”完全集成到一起。Oracle大数据机是一个硬、软件集成系统，融合了Cloudera公司的Distribution Including Apache Hadoop和Cloudera Manager，以及一个开源R。该系统采用Oracle Linux操作系统，配备有Oracle NoSQL数据库社区版本和Oracle HotSpot Java虚拟机。同时，甲骨文公司还宣布推出了最新软件产品Oracle Big Data Connectors。该产品可以帮助客户利用Oracle 数据库11g轻松整合存储在Hadoop和Oracle NoSQL数据库中的数据。

借助Oracle Exadata 数据库云服务器、Oracle Exalogic中间件云服务器与Oracle Exalytics商务智能云服务器，配备有Oracle Big Data Connectors软件的Oracle大数据机将能够满足客户在企业数据中心内获取、组织和分析大数据的所有需求。

### 3.惠普（HP）

如图1所示，大数据厂商市场份额第一是Vertica，目前已被HP收购。在大数据方面，HP的收购还包括：2010年9月3日，惠普以23.5亿美元收购了存储企业3PAR，收购之后3PAR存储业务已经连续6个季度保持100%的增长，成为增长最快的高端存储平台，同时也是惠普目前营收最大的存储产品阵列；2011年 8月惠普以100亿美元收购了英国第二大软件商Autonomy，该公司擅长基于语义计算的数据处理和数据挖掘，其软件被设计用来识别结构化数据和非结构化数据之间的关系。

惠普的大数据解决方案包括：1）HP StoreOnce全新重复数据删除解决方案，帮助企业在更短时间内保护更多数据，从而在数据爆发式增长时更好地应对风险。全新解决方案首次在单一系统中实现了高达100TB/小时的备份性能和40TB/小时的数据恢复性能；2）惠普融合云（HP Converged Cloud）、采用Autonomy Intelligent Data Operating Layer（IDOL）10的HP Data Protector 7，让企业理解并使用网络点击流量、浏览及交易数据，从而发掘新趋势、机遇及风险行动资产，从而促进业务增长及利润；3）新版惠普Vertica分析平台（HP Vertica Analytics Platform）Vertica 6，让企业能够在任何地点、使用任何接口连接、分析和管理各种类型的信息，VerticaFlexStore架构为大数据分析提供灵活的框架，与Hadoop、Autonomy或任何其它结构化、非结构化或半结构化数据源的高级集成或联合。

### 4.其他厂商和研究机构

英特尔（Intel）与麻省理工学院（MIT）成立“英特尔科学技术中心”，重点研究大数据技术，在其新一代处理器中也增加了对大数据进行处理的新技术。SAS、EMC等其他公司也分别推出各自的大数据技术解决方案。

研究机构如麦肯锡(McKinsey)，推出专门研究报告认为大数据是下一波科技竞争的前沿。世界经济论坛(WEF)则发表了“大数据，大冲击”的预测报告。联合国成立了全球脉搏（Global Pulse），致力于用大数据技术感知全球跳动的“脉搏”。

## 国内主要互联网公司的大数据战略布局与应用

随着互联网各类网络应用的不断深入，中国的大数据技术与应用的快速发展已成为不容忽视的事实。目前国内各ICT企业，特别是大型互联网企业，都开始对大数据的存储、处理和应用进行战略布局。

### 1.百度

百度作为中国最大的搜索引擎，在中国和中文互联网领域各项排行中不是最大就是最多。2012年，百度日均抓取约10亿网页，处理超过100PB（1PB=1024TB）的数据。过去10年，百度网页搜索库已从500万猛增到了500亿个页面。从公开的材料看，百度的大数据战略往往与云计算

绑定在一起，强调大数据储存与处理能力。2011年8月，百度宣布将用三年的时间建立一个全国最大的数据中心，并且主打“绿色”。通过对大数据流量的把握，百度经过设计，降低设备能耗、减少服务器、日间侧重商业业务、夜间侧重数据业务，从而让“百度的单体十万台服务器的数据中心，PUE每降低0.1，一年就可为百度节省上千万元的成本。”

### 2.腾讯

腾讯自称“目前中国最大的互联网综合服务提供商之一，也是中国服务用户最多的互联网企业之一”，拥有超过7.52亿QQ 即时通讯活跃账户，1亿微信用户、4.25亿微博用户和超过1亿的视频用户。在积累了个人用户多方面的海量数据后，2012年腾讯提出了“大数据营销”的概念。腾讯网总编辑陈菊红表示“将从这些海量数据中挖掘、分辨出用户的行为模式、兴趣偏好等，打造专属于每个人的智慧门户。” 腾讯不仅在各大产品线中都设置了数据挖掘团队，还在和一些第三方数据挖掘公司、营销公司展开合作洽谈，充分挖掘用户在网上的行为、关系、UGC(用户产生的内容)等数据，“通过合理的方法找到对企业有帮助的数据，并且将营销预算合理的分配在为数众多的数据来源平台上”，从而提高营销效率。2011年4月腾讯追加在天津的数据中心建设投资，欲建立亚洲最大的数据储备处理中心。

### 3.淘宝

相比中国用户最多的两家互联网企业，淘宝在大数据方面的举措丝毫不逊色，因为几乎所有淘宝业务都依赖淘宝数据库。每天大约有6000万用户登录淘宝网，约20亿页面浏览量（PV）。淘宝所使用的OceanBase分布式数据库，在基准数据和增量数据基础上，实现不同部门对数千亿条记录、数百TB数据上的跨行跨表事务共同完成，并支持每天4000～5000万的更新操作。早在2009年淘宝便自建大型数据库，并通过对全国淘宝购买数据的挖掘发布了2011年淘宝中国地图，对其掌握的大量用户交易数据进行了形象的展示。在利用大数据为提高用户购物体验的旗号下，淘宝根据长尾原理充分利用大数据挖掘技术，建设开放平台，提供各种增值服务。

### 4.盛大网络

盛大网络提供的文学和游戏服务吸引了为数众多的用户。2012年8月盛大调整了旗下盛大创新院的组织架构，将研究焦点放到了海量数据挖掘与智能推荐技术，深度把握个性化用户需求，将“介绍一个大家喜欢的内容”，而变为“推荐一个你会喜欢的内容”，不仅提升用户体验，而且将发展大数据作为盛大向视频和移动领域进军的机遇，将其作为未来10年赖以生存的核心竞争力予以高度重视。

### 5.中国移动

作为中国最大的移动通讯运营商，截至2012年4月底，中国移动用户数已经达到6.7亿。同时，中国移动正在谋求从移动运营商的管道角色向客户端制造和云端服务两个方向发展。而大数据业务的投入，为此提供了机遇。2011年第四季度中国移动先后与内蒙古自治区和黑龙江省签署合作协议，在呼和浩特、哈尔滨建设全国规模最大、技术最先进、能耗最低的云计算数据中心。2012年2月又确定在成都建立西部最大数据中心，完成了其在国内数据中心的三大数据基地布局。

## 国内外大数据布局与应用比较

国内互联网企业大数据的布局虽然略迟于国外，但从规模和投入上不容小觑。国内外在此领域的建设基本同步，体现在以下三个方面：第一，国外、国内大型互联网企业对大数据布局都加大投入规模，不仅是物理存储设备和处理能力的建设，也加强了分析工具的开发与分析人才队伍建设；第二，它们都在思考如何用足、用好大数据，期待从数据中挖掘潜在的巨大价值，使其为企业自身、用户和第三方带来便利与收益；第三，它们赋予大数据在数据之外的意义，都将大数据作为企业向其他领域延伸、转型的机会。

但是，限于国内外互联网发展水平、视野和其他产业积淀的不同，在大数据建设重点、建设方式和长远战略上存在差异。

在建设重点上，国内企业侧重于物理上数据存储能力建设。无论是百度、腾讯，还是淘宝、中国移动都推出了各自数据中心项目，通常以容量来衡量成就。而国外企业则已经主要侧重分析工具手段和围绕用户的解决方案开发，已经明确了大数据的盈利方式，并沿此方向不断深入。国内企业的大数据盈利更多是在探索阶段。

因为建设重点不同，国内企业往往采取“各自为战”、“平地起楼”的建设方式，从基础层面分头进行大数据存储或处理的开发。国外企业却

多采用收购兼并、合作开发多种方式来进行建设，推进大数据存储、处理、分析综合发展，而不偏于一隅。

大数据就是网络社会的未来，国外企业对大数据的提法看似“务虚”，但实际上目光长远，如IBM的智慧地球，真正体现了战略思考。而国内企业在更长时间、更广范围上的全球化布局上略显不足，大多都只以当下国内市场为目标进行大数据的定位与思考。

但在大数据具体应用上，国内企业的差异不大。首先，是为自身服务，通过大数据的开发，获取自身运行数据，为更科学、高效的组织结构安排提供条件，如百度的绿色数据中心建设。借助大数据的开发，让企业自身也变得更为灵巧，为涉足其他领域提供了机遇，如英特尔和中国移动。其次，企业通过深入挖掘用户大数据，对其行为、习惯有更为准确的把握，可以不断改善产品和服务，提升用户体验。最后，大数据的挖掘为其他商业企业营销和社会智能部门服务与管理提供依据，很可能会突破长期以来广告模式的霸主地位。

## 面向未来的大数据业务

硬件有价，数据无价，数据本身就是资产。正如麦肯锡的大数据专题报告所指出的那样，大数据已经渗透到每一个工商业组织内，将成为重要的生产要素、决策依据。大数据将产生大价值，增强企业的竞争力，将是下一波创新、竞争和提高生产力的前沿技术。

在市场方面，大数据业务将每年为美国全民医保带来3000亿的价值，全球因为个人地理信息的应用将额外产生6000亿美元的零售额，为美国创造14-19万个数据分析人员岗位等内容。另外，大数据也促使各大IT公司对信息管理专家及相关技术研发的需求，最近几年，甲骨文、IBM、微软、SAP和惠普已经在数据管理和分析上花费了超过150亿美元。目前数据挖掘及分析产业值约有1000亿美元，而且每年以10%的增长率在递增。在中国，大数据也会有比较大的发展空间，据国内有关机构估算，未来中国大数据潜在市场规模有望近2万亿。

在技术领域，一方面，大数据面临的有效存储、快速读写、实时分析等挑战，将对芯片、存储产业产生重要影响，还将催生一体化数据存储处理服务器、内存计算等市场。另一方面，因为大数据中蕴含的巨大价值，带来对数据快速处理和分析的迫切需求，将引发数据挖掘、商业智能市场的空前繁荣。

在应用方面，大数据业务可以在如下四个方面得到广泛的应用：1）快速地对突发事件的跟踪和响应，体现大数据“速度V”的特征；2）提高对危机行为变化的理解与判断，大数据“内容多样性V”的特征；3）精确绘制服务需求分布地图的能力，基于海量数据的统计分析；4）提高预测需求和供应变化的能力，综合运用大数据的快速、海量、多样数据的数据挖掘技术。

总体而言，大数据技术及业务发展，刚刚起步，展望未来，一片蓝海，但其中也隐含一些“礁石”，应引起注意和思考概括起来，可能有如下几点： 1）数据挖掘中的隐私问题；2）数据驱动的创新问题。

数据安全是数字化生活方式的隐忧之一。在现阶段人们已经生成了很多数据记录，比电话记录、上网痕迹、交易记录等，凡使用数字化工具都会留下记录。将来这些数据在各个环节打通以后，一个人的行为就无法隐藏，数据持有人可以从历史数据中推测判断出行为人的下一步动作——个人行动轨迹、行为轨迹，甚至思维轨迹。在纽约时报的一篇专题报道中，受用户赞同最多的一篇评论说，“他们（大数据持有方）想知道每一毫秒我们在做什么，找出我们的行为模式，比我们自己更了解我们，从我们已经忘记的随意点击中榨取我们每一分钱。”或许这些提法有些危言耸听，但绝非不可能，大数据技术的终究目的是使用事实统计数据，分析预测未来趋向。

另外，大数据时代，日渐发展的数据驱动式的决策模式，降低了决策风险，也扼杀了潜在的、没有数据支持的创新。在未来可能看到这样的情形，基于对人性和事物的感性认识，或者基于某种未经数据验证的理论，一个人或者组织推出某种服务，在进行决策时，由于缺少过往数据分析，没有数据支持，而无法通过惯性的数据驱动的决议形式而形成最终的商业项目。

综上，尽管大数据面临一些小问题，但仍是蓬勃发展的趋势，大数据技术是数字化发展的必然，它为人类全面、深刻地认识世界、认识自身提供了新的方式、新视角，这在此前的时代是无法办到的。我们期待大数据技术为人类文明的发展做出巨大贡献，让科技再次成为驱动社会发展的巨大动力。■

（作者单位：人民网研究院）