工业和信息化部电信研究院
China Academy of Telecommunication Research of MIIT

大 数 据 白 皮 书

（2014年）

工业和信息化部电信研究院

2014年5月

版权声明

前　言

当前，全球大数据产业日趋活跃，技术演进和应用创新加速发展，各国政府也逐渐认识到大数据在推动经济发展、改善公共服务，乃至保障国家安全方面的重大意义，纷纷积极出手推动。在全球大数据蓬勃发展的大背景下，我国也面临难得的发展机遇。如何抓住机遇，将我国拥有的数据资源转化为经济发展动力，是摆在政府和产业界面前的紧迫课题。

本白皮书首先追溯了大数据的起源，从新资源、新工具和新理念等角度探讨了大数据的概念；然后对大数据关键技术、应用、产业和政策环境等核心要素进行了分析，梳理提出了大数据技术体系和创新特点，简要描述了大数据应用及产业生态发展状况，分析了各国大数据政策实践及我国大数据发展的政策环境；最后针对我国大数据发展存在的问题，提出推动大数据应用、促进前沿技术创新与扩散、开放政府和公共数据资源、保护数据安全与个人隐私等方面的策略建议。

目 录

一、大数据概念探讨

大数据的应用和技术是在互联网快速发展中诞生的，起点可追溯到 2000 年前后。当时互联网网页爆发式增长，每天新增约 700 万个网页，到 2000 年底全球网页数达到 40 亿[1]，用户检索信息越来越不方便。谷歌等公司率先建立了覆盖数十亿网页的索引库[2]，开始提供较为精确的搜索服务，大大提升了人们使用互联网的效率，这**是大数据应用的起点**。当时搜索引擎要存储和处理的数据，不仅数量之大前所未有，而且以非结构化数据为主，传统技术无法应对。为此，谷歌提出了一套以分布式为特征的全新技术体系，即后来陆续公开的分布式文件系统(GFS，Google File System)、分布式并行计算(MapReduce)和分布式数据库(BigTable)等技术，以较低的成本实现了之前技术无法达到的规模。这些技术奠定了当前大数据技术的基础，可以认为**是大数据技术的源头**。

伴随着互联网产业的崛起，这种创新的海量数据处理技术在电子商务、定向广告、智能推荐、社交网络等方面得到应用，取得巨大的商业成功。这启发全社会开始重新审视**数据**的巨大价值，于是金融、电信等拥有大量数据的行业开始尝试这种新的理念和技术，取得初步成效。与此同时，业界也在不断对谷歌提出的技术体系进行扩展，使之能在更多的场景下使用。2011 年，麦肯锡、世界经济论坛等知名机构对这种**数据驱动的创新**进行了研究总结，随即在全世界兴起了一股大数据热潮。

虽然大数据已经成为全社会热议的话题，但到目前为止，“大数据”尚无公认的统一定义。我们认为，认识大数据，要把握“资源、

[1]来源：http://webmarketingtoday.com/articles/ad-anorexia/
[2]来源：http://www.google.com/about/company/history/，2008 年 7 月谷歌搜索引擎能检索的网页数就突破 1 万亿个。

技术、应用”三个层次。**大数据是具有体量大、结构多样、时效强等特征的数据；处理大数据需采用新型计算架构和智能算法等新技术；大数据的应用强调以新的理念应用于辅助决策、发现新的知识，更强调在线闭环的业务流程优化。**因此说，大数据不仅“大”，而且“新”，是新资源、新工具和新应用的综合体。

二、大数据关键技术

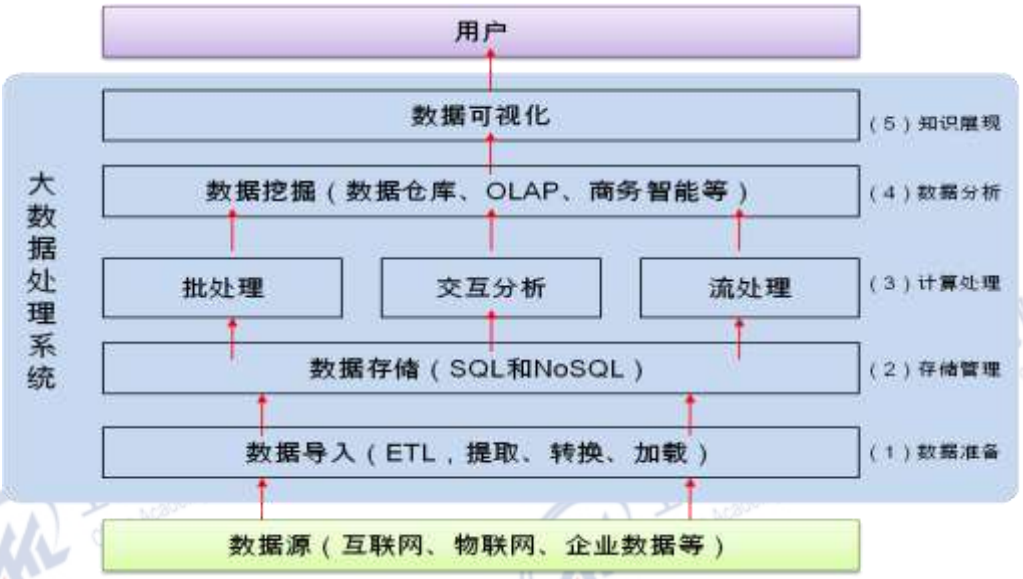
（一）大数据对传统数据处理技术体系提出挑战

大数据来源于互联网、企业系统和物联网等信息系统，经过大数据处理系统的分析挖掘，产生新的知识用以支撑决策或业务的自动智能化运转。从数据在信息系统中的生命周期看，大数据从数据源经过分析挖掘到最终获得价值一般需要经过5个主要环节，包括数据准备、数据存储与管理、计算处理、数据分析和知识展现，技术体系如图1所示。每个环节都面临不同程度的技术上的挑战。

- **数据准备环节：**在进行存储和处理之前，需要对数据进行清洗、整理，传统数据处理体系中称为ETL（Extracting，Transforming，Loading）过程。与以往数据分析相比，大数据的来源多种多样，包括企业内部数据库、互联网数据和物联网数据，不仅数量庞大、格式不一，质量也良莠不齐。这就要求数据准备环节一方面要规范格式，便于后续存储管理，另一方面要在尽可能保留原有语义的情况下去粗取精、消除噪声。
- **数据存储与管理环节：**当前全球数据量正以每年超过50%的速度增长，存储技术的成本和性能面临非常大的压力。大数据存储系统不仅需要以极低的成本存储海量数据，还要适应多样化的非结构化数据管理需求，具备数据格式上的可扩展性。
- **计算处理环节：**需要根据处理的数据类型和分析目标，采用适

当的算法模型，快速处理数据。海量数据处理要消耗大量的计算资源，对于传统单机或并行计算技术来说，速度、可扩展性和成本上都难以适应大数据计算分析的新需求。分而治之的分布式计算成为大数据的主流计算架构，但在一些特定场景下的实时性还需要大幅提升。

- **数据分析环节：**数据分析环节需要从纷繁复杂的数据中发现规律提取新的知识，是大数据价值挖掘的关键。传统数据挖掘对象多是结构化、单一对象的小数据集，挖掘更侧重根据先验知识预先人工建立模型，然后依据既定模型进行分析。对于非结构化、多源异构的大数据集的分析，往往缺乏先验知识，很难建立显式的数学模型，这就需要发展更加智能的数据挖掘技术。
- **知识展现环节：**在大数据服务于决策支撑场景下，以直观的方式将分析结果呈现给用户，是大数据分析的重要环节。如何让复杂的分析结果易于理解是主要挑战。在嵌入多业务中的闭环大数据应用中，一般是由机器根据算法直接应用分析结果而无需人工干预，这种场景下知识展现环节则不是必需的。



来源：工业和信息化部电信研究院

图 1 大数据技术框架

总的来看，大数据对数据准备环节和知识展现环节来说只是量的变化，并不需要根本性的变革。但大数据对数据分析、计算和存储三个环节影响较大，需要对技术架构和算法进行重构，是当前和未来一段时间大数据技术创新的焦点。下面简要分析上述 3 个环节面临的挑战及发展趋势。

（二）大数据存储、计算和分析技术是关键

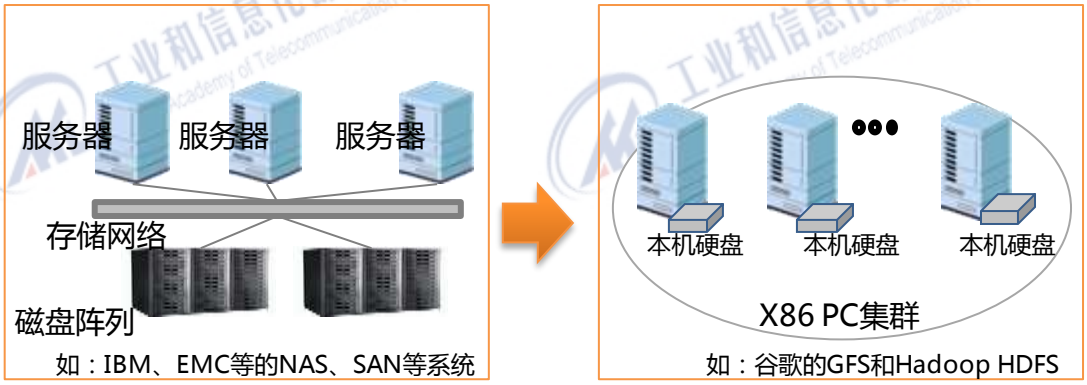
1. 大数据存储管理技术

数据的海量化和快增长特征是大数据对存储技术提出的首要挑战。这要求底层硬件架构和文件系统在性价比上要大大高于传统技术，并能够弹性扩展存储容量。但以往网络附着存储系统（NAS）和存储区域网络（SAN）等体系，存储和计算的物理设备分离，它们之间要通过网络接口连接，这导致在进行数据密集型计算（Data Intensive Computing）时 I/O 容易成为瓶颈。同时，传统的单机文件系统（如 NTFS）和网络文件系统（如 NFS）要求一个文件系统的数据必须存储在一台物理机器上，且不提供数据冗余性，可扩展性、容错能力和并发读写能力难以满足大数据需求。

谷歌文件系统（GFS）和 Hadoop 的分布式文件系统 HDFS（Hadoop Distributed File System）奠定了大数据存储技术的基础。与传统系统相比，GFS/HDFS 将计算和存储节点在物理上结合在一起，从而避免在数据密集计算中易形成的 I/O 吞吐量的制约，同时这类分布式存储系统的文件系统也采用了分布式架构，能达到较高的并发访问能力。存储架构的变化如图 2 所示。

当前随着应用范围不断扩展，GFS 和 HDFS 也面临瓶颈。虽然 GFS 和 HDFS 在大文件的追加（Append）写入和读取时能够获得很高的性能，但随机访问（random access）、海量小文件的频繁写入性能较低，

因此其适用范围受限。业界当前和下一步的研究重点主要是在硬件上基于 SSD 等新型存储介质的存储体系架构，同时对现有分布式存储的文件系统进行改进，以提高随机访问、海量小文件存取等性能。



来源：工业和信息化部电信研究院

图 2 大数据存储架构的变化

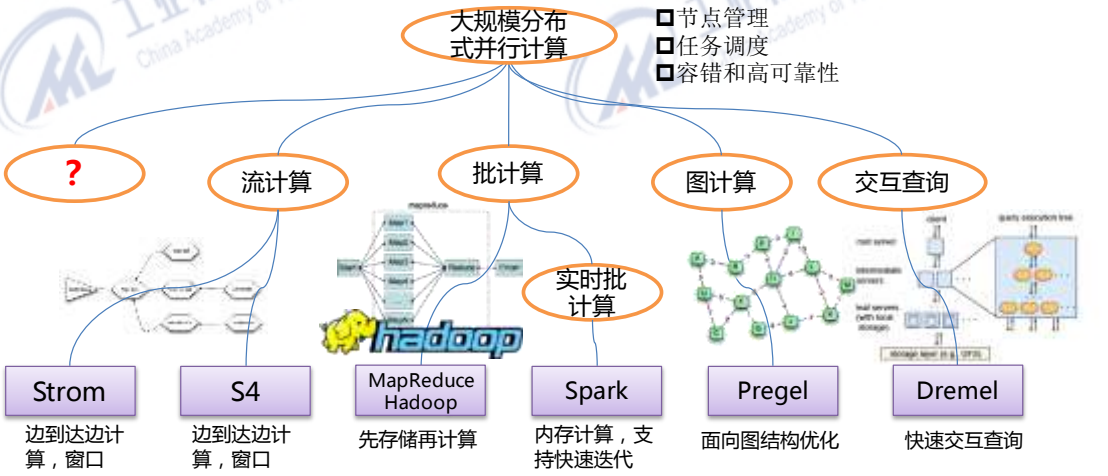
大数据对存储技术提出的另一个挑战是多种数据格式的适应能力。格式多样化是大数据的主要特征之一，这就要求大数据存储管理系统能够适应对各种非结构化数据进行高效管理的需求。数据库的一致性（Consistency）、可用性（Availability）和分区容错性（Partition-Tolerance）不可能都达到最佳，在设计存储系统时，需要在 C、A、P 三者之间做出权衡。传统关系型数据库管理系统（RDBMS）以支持事务处理为主，采用了结构化数据表的管理方式，为满足强一致性（C）要求而牺牲了可用性（A）。

为大数据设计的新型数据管理技术，如谷歌 BigTable 和 Hadoop HBase 等非关系型数据库（NoSQL，Not only SQL），通过使用“键-值（Key-Value）”对、文件等非二维表的结构，具有很好的包容性，适应了非结构化数据多样化的特点。同时，这类 NoSQL 数据库主要面向分析型业务，一致性要求可以降低，只要保证最终一致性即可，给并发性能的提升让出了空间。谷歌公司在 2012 年披露的 Spanner 数据库，通过原子钟实现全局精确时钟同步，可在全球任意位置部署，

系统规模可达到 100 万～1000 万台机器。Spanner 能够提供较强的一致性，还支持 SQL 接口，代表了数据管理技术的新方向。整体来看，未来大数据的存储管理技术将进一步把关系型数据库的操作便捷性特点和非关系型数据库灵活性的特点结合起来，研发新的融合型存储管理技术。

2. 大数据并行计算技术

大数据的分析挖掘是数据密集型计算，需要巨大的计算能力。与传统“数据简单、算法复杂”的高性能计算不同，大数据的计算是数据密集型计算，对计算单元和存储单元间的数据吞吐率要求极高，对性价比和扩展性的要求也非常高。传统依赖大型机和小型机的并行计算系统不仅成本高，数据吞吐量也难以满足大数据要求，同时靠提升单机 CPU 性能、增加内存、扩展磁盘等实现性能提升的纵向扩展（Scale Up）的方式也难以支撑平滑扩容。



来源：工业和信息化部电信研究院

图 3 针对不同计算场景发展出特定分布式计算框架

谷歌在 2004 年公开的 MapReduce 分布式并行计算技术，是新型分布式计算技术的代表。一个 MapReduce 系统由廉价的通用服务器构成，通过添加服务器节点可线性扩展系统的总处理能力（Scale Out），在成本和可扩展性上都有巨大的优势。谷歌的 MapReduce 是其内部网

页索引、广告等核心系统的基础。之后出现的开源实现 Apache Hadoop MapReduce 是谷歌 MapReduce 的开源实现，目前已经成为目前应用最广泛的大数据计算软件平台。

MapReduce 架构能够满足“先存储后处理”的离线批量计算(batch processing）需求，但也存在局限性，最大的问题是时延过大，难以适用于机器学习迭代、流处理等实时计算任务，也不适合针对大规模图数据等特定数据结构的快速运算。

为此，业界在 MapReduce 基础上，提出了多种不同的并行计算技术路线，如图 3 所示。如 Yahoo 提出的 S4 系统、Twitter 的 Storm 系统是针对“边到达边计算”的实时流计算(Real time streaming process)框架，可在一个时间窗口上对数据流进行在线实时分析，已经在实时广告、微博等系统中得到应用。谷歌 2010 年公布的 Dremel 系统，是一种交互分析（Interactive Analysis）引擎，几秒钟就可完成 PB（1PB=10^{15}B）级数据查询操作。此外，还出现了将 MapReduce 内存化以提高实时性的 Spark 框架、针对大规模图数据进行了优化的 Pregel 系统等等。

来源：www.hortonworks.com

图 4 Hadoop 2.0 将资源管理和任务调度分离

针对不同计算场景建立和维护不同计算平台的做法，硬件资源难以复用，管理运维也很不方便，研发适合多种计算模型的通用架构成为业界的普遍诉求。为此，Apache Hadoop 社区在 2013 年 10 月发布的 Hadoop 2.0 中推出了新一代的 MapReduce 架构。新架构的主要变

化是将旧版本 MapReduce 中的任务调度和资源管理功能分离，形成一层与任务无关的资源管理层（YARN）。如图 4 所示，YARN 对下负责物理资源的统一管理，对上可支持批处理、流处理、图计算等不同模型，为统一大数据平台的建立提供了新平台。基于新的统一资源管理层开发适应特定应用的计算模型，仍将是未来大数据计算技术发展的重点。

3. 大数据分析技术

在人类全部数字化数据中，仅有非常小的一部分（约占总数据量的 1%）数值型数据得到了深入分析和挖掘（如回归、分类、聚类），大型互联网企业对网页索引、社交数据等半结构化数据进行了浅层分析（如排序）。占总量近 60%的语音、图片、视频等非结构化数据还难以进行有效的分析。

大数据分析技术的发展需要在两个方面取得突破，一是对体量庞大的结构化和半结构化数据进行高效率的深度分析，挖掘隐性知识，如从自然语言构成的文本网页中理解和识别语义、情感、意图等；二是对非结构化数据进行分析，将海量复杂多源的语音、图像和视频数据转化为机器可识别的、具有明确语义的信息，进而从中提取有用的知识。

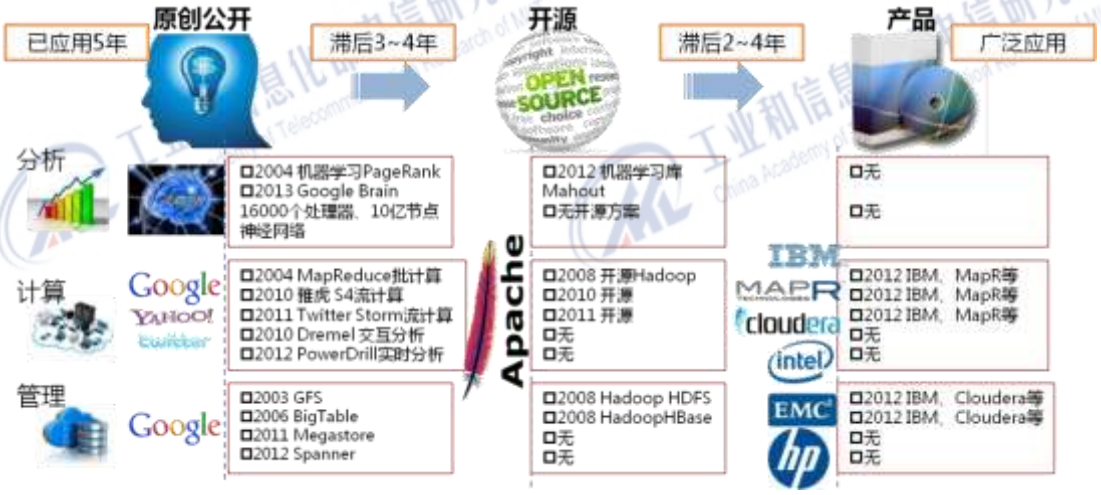
目前的大数据分析主要有两条技术路线，一是凭借先验知识人工建立数学模型来分析数据，二是通过建立人工智能系统，使用大量样本数据进行训练，让机器代替人工获得从数据中提取知识的能力。由于占大数据主要部分的非结构化数据，往往模式不明且多变，因此难以靠人工建立数学模型去挖掘深藏其中的知识。

通过人工智能和机器学习技术分析大数据，被业界认为具有很好的前景。2006 年谷歌等公司的科学家根据人脑认知过程的分层特性，提出增加人工神经网络层数和神经元节点数量，加大机器学习的规模，

构建深度神经网络，可提高训练效果，并在后续试验中得到证实。这一事件引起工业界和学术界高度关注，使得神经网络技术重新成为数据分析技术的热点。目前，基于深度神经网络的机器学习技术已经在语音识别和图像识别方面取得了很好的效果。但未来深度学习要在大数据分析上广泛应用，还有大量理论和工程问题需要解决，主要包括模型的迁移适应能力，以及超大规模神经网络的工程实现等。

（三）大数据技术创新呈现“原创-开源-产品化”的阶梯格局

当前，国际上大数据技术创新方面形成了独特的“互联网公司原创——开源扩散——IT 厂商产品化——其他企业使用”特点，如图 5 所示。



来源：工业和信息化部电信研究院

图 5 大数据技术创新格局

总结互联网企业在大数据技术创新方面的经验，如下几个方面值得关注：

第一，丰富的数据和强大的平台是大数据创新的基础条件。以谷歌为例，它的数据资源极为丰富，拥有全球网页索引库，掌握几十亿用户的搜索行为数据，建立了高分辨率的谷歌地图数据数据库，拥有

谷歌+社交数据和 YouTube 海量视频数据。谷歌的基础设施也十分强大，在全球拥有 36 个超大型数据中心，上百万台服务器。

第二，大数据的应用效益不是飞跃突进的，必须依靠长期的不断累积。从搜索、广告和推荐等成熟应用来看，大数据的应用效果并非立竿见影，其巨大的效益是在日积月累的微小进步中逐渐形成的。

第三，累积效益的获取，主要靠持续不断的技术迭代。互联网企业一直奉行敏捷开发、快速迭代的软件开发理念，往往在一两周内就能完成一个“规划、开发、测试、发布”的迭代周期。大型互联网企业通过这种长期持续“小步快跑”的研发方式，支撑了大数据应用效果的持续提升，建立了技术上的领先优势。

第四，技术和应用一体化组织，是快速迭代的保障。互联网企业之所以能够保持高效率的持续技术演进，其研发和应用一体化的组织方式是很重要的因素。与传统行业“应用者——解决方案提供商”分离的组织形态不同，互联网企业省去了解决方案供应商环节，可以迅速将需求转化为解决方案。谷歌、百度等大型互联网企业的研发人员占比一般都在 50%～70%，远远高于其他类型的公司，这为技术开发提供了强大的后盾。

最后，大数据技术发展与开源运动的结合也成为大数据技术创新中的一个鲜明特点。领先企业进行前沿创新，创新成果通过开源得到不断完善并向全社会辐射，原创与开源相得益彰，在国际上形成了一套高效运转的研发产业化体系。开源模式让人们“不必重复发明轮子”，能够降低研发和采购成本，还能够启发新的创意，加快再创新步伐。特别是开源 Apache Hadoop 的大范围应用，大大加速了大数据应用进程，一大批互联网公司和传统 IT 企业都从这种技术扩散体系中受益。在此背景下，国内大数据技术研发也应该把自主创新和开源结合起来，

以更加开放的心态融入到国际大数据技术创新潮流中去。

三、大数据应用发展

大数据的价值体现在大数据的应用上，人们关心大数据，最终是关心大数据的应用，关心如何从业务和应用出发让大数据真正实现其所蕴含的价值，从而为我们的生产生活带来有益的改变。对“大数据应用”，不同行业和不同应用者理解不同。本白皮书将大数据应用界定为：“利用分布式并行计算、人工智能等技术对海量异构数据进行计算、分析和挖掘，并将由此产生的信息和知识应用于实际的生产、管理、经营和研究中”。

整体而言，全球的大数据应用处于发展初期，中国大数据应用才刚刚起步。目前，大数据应用在各行各业的发展呈现“阶梯式”格局：互联网行业是大数据应用的领跑者，金融、零售、电信、公共管理、医疗卫生等领域积极尝试大数据。

（一） 互联网行业是大数据应用的领跑者

互联网是大数据应用的发源地，大型互联网企业是当前大数据应用的领跑者。搜索引擎作为最早的互联网大数据应用，其不断的发展推动谷歌在 2000 年左右提出了 MapReduce/BigTable 等技术，从此开启了大数据技术的新篇章。经过十多年的发展，互联网上形成了多种相对成熟的大数据应用模式，按照用途可分为以下三类：

第一，商业大数据应用，即主要以盈利为目的的大数据应用。目前常见的应用有：一是基于用户个人信息、行为、位置、微博等数据而进行的个性化推荐、交叉推荐、品牌监测等营销类大数据应用。由于其商业模式清晰可见，市场需求广泛旺盛，因此这是目前互联网上最热门最普遍的应用，被互联网广告、电子商务、微博、视频、相亲等公司普遍采用。二是基于用户、商铺的交易数据而进行的经营分析

报告、反欺诈、反虚假交易、促销和团购选品、产业集聚判断等交易辅助类大数据应用，这些应用目前已经逐渐成为电子商务企业的必备工具。三是利用网站动态数据对网络状态进行实时监控预警、网站分析优化和网络信息安全保护(如用户信息保护、防网络攻击、防钓鱼、防垃圾邮件等）的网络安全大数据应用。

第二，公共服务类大数据应用，即不以盈利为目的、侧重于为社会公众提供服务的大数据应用。典型案例如谷歌开发的流感、登革热等流行病预测应用能够比官方机构提前一周发现疫情爆发状况。国内也有搜索引擎公司提供诸如春运客流分析、失踪儿童搜寻的公益大数据服务。

第三，技术研发类大数据应用，即利用大数据技术促进前沿技术研发、持续改进产品性能的应用。互联网应用在新版本的研发中，常常进行 A/B 测试就是大数据在产品开发中的典型应用。A/B 测试中，服务商同时收集新老版本下的用户行为数据(如点击行为、访问时长、鼠标停留等)，并进行分析比对，用以指导产品后续的改进方向。另外，利用各种语言版本的网页数据不断提高翻译质量的机器翻译、利用更多话音指令不断提升质量的话音识别技术，以及无人汽车等前沿技术的研发也广泛应用了大数据技术。

（二）　大数据应用加速向传统领域拓展

大数据应用起源于互联网，正在向以数据生产、流通和利用为核心的各个产业渗透。目前金融、零售、电信、公共管理、医疗卫生等领域在积极地探索和布局大数据应用，主要呈现两种发展方向：

一是积极整合行业和机构内部的各种数据源，通过对整合后的数据进行挖掘分析，从而发展大数据应用。例如，一些新兴的大型百货商场利用大数据平台整合 POS(point of sale)机、企业 CRM(Customer

Relationship Management）系统、免费无线网络、客流监控设备等数据，对用户进行聚类分析，支撑包括商品位置摆放、打折信息投放、移动端营销、客户习惯查询、客户群路径分析等应用，提高商场营销效率和营业额。基于大数据的智慧城市决策系统也是大数据应用的重要领域，可整合来自经济、统计、民政、教育、卫生、人力等政府部门内部数据和来自物联网、移动互联网等网络数据，设计经济社会运行分析模型，支撑智慧人口、智慧医疗、智慧教育、智能物流、智能环保等相关决策应用。

二是积极借助外部数据，主要是互联网数据，来实现相关应用。例如，金融机构通过收集互联网用户的微博数据、社交数据、历史交易数据来评估用户的信用等级；证券分析机构通过整合新闻、股票论坛、公司公告、行业研究报告、交易数据、行情数据、报单数据等，试图分析和挖掘各种事件和因素对股市和股票价格走向的影响；监管机构将社交数据、网络新闻数据、网页数据等与监管机构的数据库对接，通过比对结果进行风险提示，提醒监管机构及时采取行动；零售企业通过互联网用户数据分析商品销售趋势、用户偏好等等。

从目前发展的情况来看，金融、零售和公共管理领域开展大数据应用时，两个方向都有所涉足，而电信和医疗卫生等领域更关注第一个发展方向。

（三） 大数据应用呈现初级发展阶段特征

当前大数据还未形成普遍应用的局面，对大多数企业，特别是传统领域的企业而言，还未找到有效的应用模式，总体上看呈现以下几个方面的特点：

首先，理念的应用快于数据的应用。对大数据的广泛讨论，使人们普遍认识到，数据是有价值的，人们可以通过各种方法和技术，对

数据进行分析和挖掘，从中获得对我们生产生活有利的信息和知识；任何数据都可能是有价值的，关键是看谁使用它、怎么使用它；简而言之，数据是资产。这一轮大数据浪潮，使得大数据理念迅速普及，尽管很多数据尚没有找到合适的用途，但很多公司已经将其作为资产，对其数据进行规划、存储，或自行对其开发，或者积极寻找买家或者合作者来对其进行开发。电信运营商最有可能成为典型的数据资产运营者。电信运营商掌握丰富的用户身份数据、语音数据、视频数据、流量数据和位置数据，数据的海量性、多元性和实时性使其具有经营大数据的先天优势，目前主要的电信运营商都已积极探索开发其内部大数据资源。但从目前的应用发展看，电信运营商的大数据仍主要用于支持内部的客户流失分析、营销分析和网络优化分析等，对外的应用模式尚未成型。

其次，大数据应用呈散发状，并没有形成燎原之势，目前主要集中于互联网的市场营销场景。尽管金融、电信、零售、制造、医疗、交通、物流、IT 等行业对大数据应用表现出极大热情，但目前在媒体和各种论坛上所公开的大数据应用案例仍然非常零散，这表明大家虽然都很关注大数据，但推进实际的应用仍然存在一定的困难。唯一众多企业都推出或者采纳大数据应用的领域是基于互联网的市场营销，在这一领域应用了大数据的公司不仅包括大型的互联网公司、众多专业性的中小型互联网公司，线下企业也在与互联网公司合作，积极开发这一领域的价值。

从数据源看，大数据的应用还处于自给自足的“小农经济”时代，现有的应用仍然以机构内部数据为主。以机构内部数据为主的主要原因是数据的开放和交易尚未形成市场的主流形态。以国内主要的电子商务交易平台为例，目前推出了很多大数据应用，但这些应用基本都

是为内部服务的，由于法律和数据交易机制的不健全，这些交易平台在对外开放和交易数据上仍然持谨慎态度。Gartner 的一项调查[3]显示，即使在全球，以内部数据为主仍然是大数据应用的主要特征，各行业应用最多的仍然是企业内部的交易数据（应用比例普遍超过 50%，多数行业应用比例超过 80%）和日志数据。

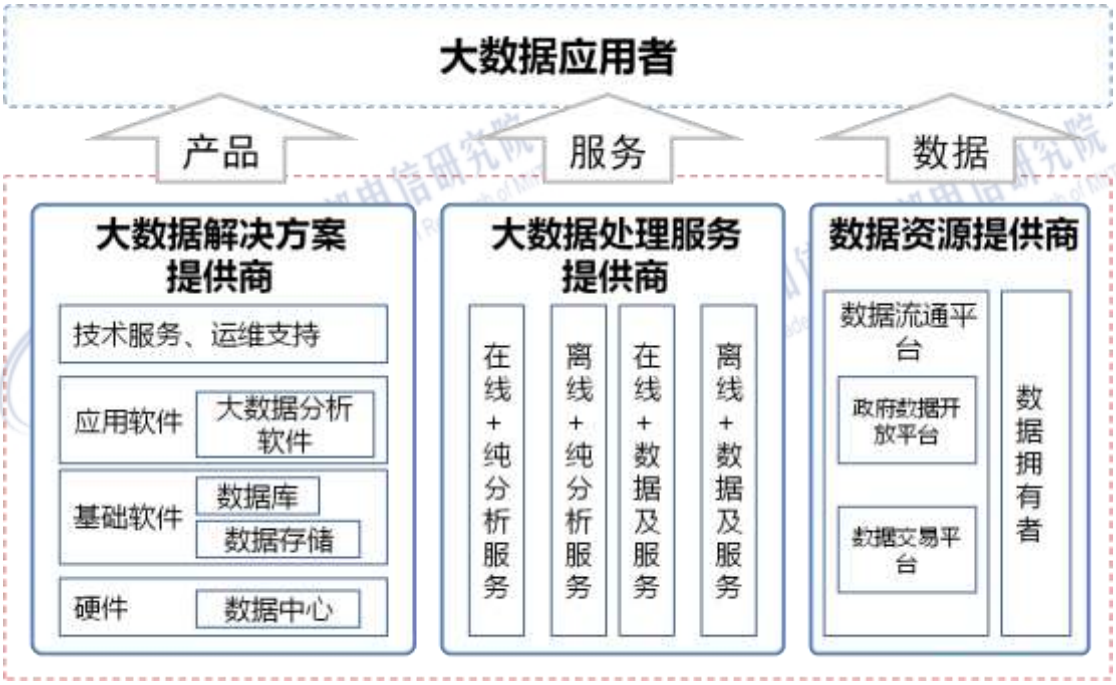
从技术角度看，大数据仍以初级应用为主，多数应用仍然使用传统分析流程和工具，只是扩大了数据的来源、增加了数量。调研发现，与传统数据分析相比，新的大数据应用虽然开始使用非结构化数据，但在实际应用过程中，这些非结构化数据只是被压缩、清洗和结构化后，放入传统的 ETL 和分析流程中去。另一些大数据应用通过采用云存储和云处理技术，提高了数据处理效率，从而增加了数据处理的规模，但这些应用也仍然采用原有的 ETL 和分析流程。缺乏应用模式上的创新，使得目前大数据应用仍停留在初级技术阶段。

从应用效果看，目前的大数据应用以延续改善现有业务和产品为主，突破性创新应用尚不多见。以最常见的互联网营销大数据应用为例，在大数据兴起之前，精准营销和个性化推荐一直是企业营销活动的追求方向，新兴数据源和大数据技术的兴起使得企业进一步改善其营销技能，使其精准营销能力进一步增强，这是对企业旧有营销能力的改善。目前大家议论比较多的突破性创新如网上小贷业务，这项业务完全改变了过去金融机构贷款的流程、信用评价和控制风险的方式，从而极大地降低了贷款的成本、扩大了贷款的范围。但目前这样的突破性创新并不多见。Gartner 的调查显示[4]，企业投资大数据的主要目的在于改善客户服务、流程优化、精准营销和削减成本等，而新产品/新商业模式这种突破性创新的方向并不是企业的主要目的。

[3]来源：Gartner，“调查分析：大肆宣传下的 2013 年大数据应用程序”，2013 年 9 月
[4]来源：Gartner，“调查分析：大肆宣传下的 2013 年大数据应用程序”，2013 年 9 月

四、大数据产业生态

随着大数据技术不断演进和应用持续深化，以数据为核心的大数据产业生态正在加速构建。从实践情况看，**大数据产业生态中主要包括大数据解决方案提供商、大数据处理服务提供商和数据资源提供商三个角色，分别向大数据的应用者提供大数据服务、解决方案和数据资源。**大数据产业生态构成如图 6 所示。当前大数据产业还处于构建的初期，呈现规模很小、增速快的特点，据 Wikibon 公司的报告[5]，2013 年全球大数据市场总体规模为 181 亿美元，年度增幅达 61%，预计到 2017 年还将维持 30%的年增速。



来源：工业和信息化部电信研究院

图 6 大数据产业生态示意图

（一） 传统 IT 厂商加快向大数据解决方案提供商转型

大数据解决方案提供商面向企业用户提供大数据一站式部署方案，覆盖数据中心和服务器等硬件、数据存储和数据库等基础软件、大数据分析应用软件以及技术运维支持等方面内容。其中，大数据基

[5]来源：Wikibon 公司“2012-2017 年大数据企业收入和市场预测”，2013 年 4 月。

础软件和应用软件是大数据解决方案中的重点内容。当前，企业提供的大数据解决方案大多基于 Hadoop 开源项目，例如，IBM 基于 Hadoop 开发的大数据分析产品 BigInsights、甲骨文融合了 Hadoop 开源技术的大数据一体机、Cloudera 的 Hadoop 商业版等。

大数据解决方案提供商中，主要包括传统 IT 厂商和新兴的大数据创业公司。传统 IT 厂商主要有 IBM、HP 等解决方案提供商以及甲骨文、Teradata 等数据分析软件商。它们大多以原有 IT 解决方案为基础，融合 Hadoop，形成融合了结构化和非结构化两条体系的“双栈”方案。通过一系列收购来提升大数据解决方案服务能力，成为这些 IT 巨头的主要策略。

国际上也诞生了一批专门提供非结构化数据处理方案的新兴创业公司。这些公司包括 Cloudera、Hortonworks、MapR 等，它们主要基于 Hadoop 开源项目，开发 Hadoop 商业版本和基于 Hadoop 的大数据分析工具，单独或者与传统 IT 厂商合作提供企业级大数据解决方案。这些新兴大数据企业成为资本市场的热点。

国内华为、联想、浪潮、曙光等一批 IT 厂商也都纷纷推出大数据解决方案。但总体上，国内大数据解决方案提供商实力较弱，产品一些关键行业还未形成影响力，新兴大数据解决方案初创企业也凤毛麟角。

（二） 云服务商成为大数据处理服务提供商的主体

大数据处理服务提供商主要以服务的方式为企业和个人用户提供大数据海量数据分析能力和大数据价值挖掘服务。按照服务模式进行划分，大数据处理服务提供商可以分为以下四类。

第一类是在线纯分析服务提供商。此类服务商主要是互联网企业、大数据分析软件商和新创企业等，通过 SaaS 或 PaaS 云服务形式为用

户提供服务。典型的服务如谷歌提供的大数据分析工具 Big Query、亚马逊提供的云数据仓库服务 RedShift、微软的 Azure HDInsigh1010data 提供的商业智能服务等。国内一些云服务商也逐步开始提供大数据相关云服务，如阿里云的开放数据处理服务（ODPS）、百度的大数据引擎、腾讯的数据云等。

第二类是既提供数据又提供分析服务的在线提供商。此类服务商主要是拥有海量用户数据的大型互联网企业，主要以 SaaS 形式为用户提供大数据服务，服务背后以自有大数据资源为支撑。典型的服务如如谷歌 Facebook 的的自助式广告下单服务系统、Twitter 基于实时搜索数据的产品满意度分析等。国内百度推出的大数据营销服务“司南”就属于此类。

第三类是单纯提供离线分析服务的提供商。此类服务商主要为企业提供专业、定制化的大数据咨询服务和技术支持，主要集中为大数据咨询公司、软件商等，例如专注于大数据分析的奥浦诺管理咨询公司（Opera Solutions）、数据分析服务提供商美优管理顾问公司（Mu Sigma）等。

第四类是既提供数据又提供离线分析服务的提供商。此类服务商主要集中在信息化水平较高、数据较为丰富的传统行业。例如日本日立集团（Hitachi）于 2013 年 6 月初成立的日立创新分析全球中心，其广泛收集汽车行驶记录、零售业购买动向、患者医疗数据、矿山维护数据和资源价格动向等庞大数据信息，并基于收集的海量信息开展大数据分析业务。又如美国征信机构 Equifax 基于全球 8000 亿条企业和消费者行为数据，提供 70 余项面向金融的大数据分析离线服务。

（三） 大数据资源提供商应运而生

既然数据成为了重要的资源和生产要素，必然会产生供应与流通

需求。数据资源提供商因此应运而生，它是大数据产业的特有环节，也是大数据资源化的必然产物。数据资源提供商，包括数据拥有者和数据流通平台两个主要类型。

数据拥有者可以是企业、公共机构或者个人。数据拥有者通常直接以免费或有偿的方式为其它有需求的企业和用户提供原数据或者处理过的数据。例如美国电信运营商 Verizon 推出的大数据应用精准营销洞察（Precision Market Insights），将向第三方企业和机构出售其匿名化和整合处理后的用户数据。国内阿里巴巴公司推出的淘宝量子恒道、数据魔方和阿里数据超市等，属于此种类型。

数据数据流通平台是多家数据拥有者和数据需求方进行数据交换流通的场所。按平台服务目的不同，可分为政府数据开放平台和数据交易市场：

- 政府数据开放平台：主要提供政府和公共机构的非涉密数据开放服务，属于公益性质。目前全球不少国家已经加入到开放政府数据行动，推出公共数据库开放网站，例如美国数据开放网站 Data.gov 目前已有超过 37 万个数据集、1209 个数据工具、309 个网页应用和 137 个移动应用，数据源来自 171 个机构。国内地方政府数据开放平台开始出现，如国家统计局的国家数据网站、北京市政府和上海市政府的信息资源平台等数据开放平台正在建设过程中。
- 数据交易市场：商业化的数据交易活动催生了多方参与的第三方数据交易市场。国际上目前比较有影响力的有微软的 Azure Data Marketplace、被甲骨文收购的 BlueKai、DataMarket、Factual、Infochimps、DataSift 等等，主要提供地理空间、营销数据和社交数据的交易服务。大数据交易市场发展刚刚起

步，在市场机制、交易规则、定价机制、转售控制和隐私保护等方面还有很多工作要做。国内，2014 年 2 月，在北京市和中关村管委会指导下，中关村大数据交易产业联盟成立，将在国内推动国内大数据交易相关规范化方面开展工作。

五、大数据政策环境

（一） 国外政府从数据、技术和应用三方面推进大数据发展

美英日澳等国家政府高度重视大数据产业发展，自 2012 年来密集出台多项专门政策予以支持。从各国举措来看，政策着力点主要在于三个方面。其一是开放数据，给予产业界高质量的数据资源；其二是在前沿及共性基础技术上增加研发投入；其三是积极推动政府和公共部门应用大数据技术。

美国在推动大数据研发和应用上最为迅速和积极，强化顶层设计，力图引领全球大数据发展。2012 年美国联邦政府就在全球率先推出“大数据行动计划（Big data initiative）”[6]，重点在基础技术研究和公共部门应用上加大投入。在该计划支持下，加州大学伯克利分校开发了完整的大数据开源软件平台“伯克利数据分析软件栈（Berkeley Data Analytics Stack）[7]，其中的内存计算软件 Spark 的性能比 Hadoop 提高近百倍[8]，对产业界大数据技术走向产生巨大影响。美国政府还在积极推动数据公开，已开放 37 万个数据集和 1209 个数据工具，并在 2013 年 5 月初进一步要求，政府必须实现新增和经处理数据的开放和机器可读，激发大数据创新活力。同时，美国政府也是大数据的积极使用者，2013 年曝光的棱镜门事件显示出美国

[6]来源：http://www.whitehouse.gov/sites/default/files/microsites/ostp/big_data_press_release_final_2.pdf
[7]来源：https://amplab.cs.berkeley.edu/software/
[8]来源：http://spark.apache.org/

国家安全部门大数据应用的强大实力，其应用范围之广、水平之高、规模之大都远远超过人们的想象。2012-2013 年间美国国家安全局（NSA）、联邦调查局（FBI）及中央情报局（CIA）等联邦政府机构还大量采购亚马逊的云服务，以支撑其大数据应用。随着应用的深入，美国政府对大数据带来的负面影响也更加重视，白宫 2014 年 5 月发布的《大数据：抓住机遇，守护价值》报告中提醒，在发挥正面价值的同时，应该警惕对大数据应用对隐私、公平等长远价值带来的负面影响[9]。

英国将大数据列为战略性技术，给予高度关注。英国政府紧随美国之后，推出一系列支持大数据发展举措。首先是给予研发资金支持。2013 年 1 月，英国政府向航天、医药等 8 类高新技术领域注资 6 亿英镑研发，其中大数据技术获得 1.89 亿英镑的资金[10]，是获得资金最多的领域。其次是促进政府和公共领域的大数据应用。据测算，通过合理、高效使用大数据技术，英国政府每年可节省约 330 亿英镑，相当于英国每人每年节省约 500 英镑。为了在医疗领域更好的应用大数据，2013 年 5 月，英国政府和李嘉诚基金会联合投资设立全球首个综合运用大数据技术的医药卫生科研机构，将透过高通量生物数据，与业界共同界定药物标靶，处理目前在新药开发过程中关键的瓶颈，之后还将汇集遗传学、流行病学、临床、化学和计算机科学等领域的顶尖人才，集中分析庞大的医疗数据。

日本政府把大数据作为提升日本竞争力的关键。日本政府认为，提升日本竞争力，大数据应用不可或缺。日本在新一轮 IT 振兴计划中把发展大数据作为国家战略的重要内容，新的 ICT 战略重点关注大数据应用技术。日本总务省 2012 年 7 月推出了新的综合战略“活力

[9]来源：http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf
[10]来源：http://www.e-gov.org.cn/xinxihua/news003/201305/141545.html

ICT 日本”，将重点关注大数据应用，并将其作为 2013 年六个主要任务之一，聚焦大数据应用所需的、社会化媒体等智能技术开发，以及在新医疗技术开发、缓解交通拥堵等公共领域的应用。

此外，澳大利亚、新加坡等国也非常重视大数据发展。2013 年 8 月初，澳大利亚出台公共服务大数据政策，提出了大数据分析的实践指南，希望通过大数据分析系统提升公共服务质量，增加服务种类，为公共服务提供更好的政策指导。在新加坡政府，多个国际领先企业在当地设立大数据技术研发中心，加速数据分析技术的商业应用。2014 年初，新加坡资讯通信发展管理局(IDA)还聘请了首任首席数据科学家，专门推进政府数据的开放和价值开发。

（二） 我国大数据发展环境持续完善

我国大数据发展的宏观政策环境不断完善。2012 年以来，科技部、发改委、工信部等部委在科技和产业化专项陆续支持了一批大数据相关项目，在推进技术研发方面取得了积极效果。2013 年 6 月工信部发布的《电信和互联网用户个人信息保护规定》，根据《全国人民代表大会常务委员会关于加强网络信息保护的决定》，进一步界定了个人信息的范围，提出了个人信息的收集和使用规则、安全保障等要求，为大数据应用中的个人信息保护设立了法律法规屏障。2014 年《政府工作报告》明确提出，“以创新支撑和引领经济结构优化升级；设立新兴产业创业创新平台”，在新一代移动通信、集成电路、大数据等方面赶超先进，引领未来产业发展。

地方政府积极推动大数据发展，2013 年以来陆续出台了推进计划。总体上看各地大数据发展政策各有侧重，形成了不同的模式。**模式一是强调研发及公共领域应用。**如上海市《推进大数据研究与发展三年行动计划》提出，将在三年内选取医疗卫生、食品安全、终身教

育、智慧交通、公共安全、科技服务6个有基础的领域，建设大数据公共服务平台。**模式二是强调以大数据引领产业转型升级。**如北京中关村《关于加快培育大数据产业集群 推动产业转型升级的意见》提出，要充分发挥大数据在工业化与信息化深度融合中的关键作用，推动中关村国家自主创新示范区产业转型升级。**模式三是强调建立大数据基地，吸纳企业落户。**如重庆、贵州、陕西、湖北等地都提出建设大数据产业基地的计划，力图将大数据培育成本地的支柱产业。在地方积极推动大数据发展的同时，也应警惕将“大数据”简单等同于“大数据中心”、盲目上马大规模园区建设的潜在过热风险。

六、问题与政策思考

总体上，我国的大数据产业具备良好基础，发展前景广阔。一是一批世界级的互联网公司在大数据应用上不断推陈出新，智能搜索、广告、电商、社交等借助大数据技术持续进化，互联网金融、O2O（online to offline）等应用借助大数据向线下延伸。二是大数据技术紧跟国际先进水平，具备建设和运营世界最大规模大数据平台的能力，单集群规模达5000到10000台服务器，数据管理规模达到EB（$1EB=10^{18}B$）级别，在机器学习等方面也有所突破。三是当前和未来一段时间，我国面临着经济结构转型升级、政府和公共服务改进提升等紧迫任务，这些方面大数据都有广阔的应用前景。

（一） 我国大数据发展面临的问题

应该认识到，大数据在全球的发展还都处于初期，技术、制度、观念等方面都需要改变。对我国来说，数据资源不丰富、技术差距大和法律法规不完善是当前大数据发展面临的主要问题。

一是我国数据源不够丰富，数据开放程度较低。丰富的高质量数据资源是大数据产业发展的前提。近几年在互联网产业及金融、电信

信息化快速发展的带动下，我国数据资源总量有了快速增长，已达到全球的13%，但其他行业受信息化水平制约，数据储量仍不丰富。已有数据资源还存在标准化、准确性、完整性低，利用价值不高的情况。同时，我国政府、企业和行业信息化系统建设中受到各种因素制约，形成了众多“信息孤岛”，数据开放程度严重滞后。建立良性发展的数据资源储备与共享体系，是我国大数据发展的首要问题。

二是我国大数据技术存在水平不高，技术扩散不畅的问题。我国大数据技术的发展模式也与全球类似，互联网企业具备快速将国际先进的开源大数据技术整合到自身系统中的能力，并构建了单集群上万节点的大型系统，但仍缺乏原创技术，对开源社区的贡献不足，进而对前沿技术路线的影响比较微弱。同时，由于本土开源社区等产业组织发育滞后，国内领先企业在大数据方面的技术创新也难以向社会扩散。

三是大数据相关的法律法规有待进一步完善。随着大数据挖掘分析将越来越精准、应用领域不断扩展，个人隐私保护和数据安全变得非常紧迫。在隐私保护方面，现有的法律体系面临着两个方面的挑战：一是法律保护的个人隐私主，要体现为“个人可识别信息（PII，Personally identifiable information”，但随着技术的推进，以往并非PII的数据也可能会成为PII，使得保护范围变得模糊。二是以往建立在“目的明确、事先同意、使用限制”等原则之上的个人信息保护制度，在大数据场景下变得越来越难以操作。而我国个人信息保护、数据跨境流动等方面的法律法规尚不健全，这成为制约大数据产业健康发展的重要原因之一。需要结合我国法治建设的实际情况，探索通过行业自律等方式弥补法律体系不完善的弊端。

（二） 推动我国大数据产业发展的对策思考

在总体思路上，需要首先明确我国大数据发展的战略目标和战略重点，统筹谋划大数据应用、关键技术研发与产业培育、数据开放与数据保护、市场监管、法律法规等关键布局，引导国内各地大数据发展方向，避免一哄而上的盲目发展。

在大数据应用上，一是政务和公共服务领域的应用，重点面向改善民生服务和城市治理等方面，积极推动环保、医疗、教育、交通等关键领域的大数据整合与集成应用，进一步提高政务和公共服务效率。二是市场化应用方面，重点在跨行业的大数据应用方面出台推动政策，促进互联网、电信、金融等企业与其他行业开展大数据融合与应用创新，带动全社会大数据应用不断深化。

在技术创新上，一是要加强大数据技术研发方向的前瞻性和系统性，近期重点支持深度学习与人工智能、实时大数据处理、海量数据存储管理、交互式数据可视化和应用相关的分析技术。二是要聚集产学研用力量形成合力，力争在大数据平台级软件上实现突破，以此为核心发展开源生态。三是创新科研项目支持方式，将开源和开放标准作为考核指标，通过直接补助或后补助方式激励企业和科研机构参与开源技术发展，促进大数据技术扩散。

在政府数据开放上，建议推进政府和公用事业领域数据资源的普查工作，并按照相关法规制定政府和公共数据开放中的安全和隐私保护检查表，对可能涉及国家安全和公民隐私的风险点进行严格控制。在此基础上，按敏感性对政府和公共数据进行分类，确定开放优先级，制定分步骤的数据开放路线图。同时，政府也应积极规范和引导商业化的大数据交易活动，为数据资源的流通创造有利条件。

在个人信息保护上，国际上一些机构提出，为了释放大数据潜力，

监管的重点应该“从数据收集环节，转移数据使用环节”。我们要密切关注国际上立法理念的演变趋势，结合技术发展趋势和我国国情对相关制度进行前瞻性研究。同时，为了解决当前个人信息和数据保护的紧迫需求，可依托行业组织及时总结业界的最佳实践，逐步形成行业共识，在试点成熟后上升为标准或法律法规并推动实施，为大数据的健康发展保驾护航。

工业和信息化部电信研究院
地　　址：北京市海淀区花园北路 52 号
邮政编码：100191
联系电话：010-62303621、62301204
传　　真：010-62304980