大数据标准化白皮书

指导单位：工业和信息化部软件服务业司

国家标准化管理委员会工业二部

编制单位：中国电子技术标准化研究院

二零一四年六月

版权：

使用声明：

目 录

1.前言

1.1 研究背景

大数据[1]（Big Data）是一场革命，将改变我们的生活、工作和思维方式。继移动互联网、云计算后，大数据逐渐成为对于 ICT 产业具有深远影响的技术变革。大数据技术的发展与应用，将对社会的组织结构、国家的治理模式、企业的决策架构、商业的业务策略以及个人的生活方式产生深刻影响。

我国正处于工业化向信息化发展的转型时期，信息的公开、共享与服务成为时代发展的主题。信息逐渐成为与物质和能源同等重要的资源，以开发和利用信息资源为目的的经济活动迅速扩大，逐渐占据或超越工业活动在国民经济活动中的地位。大数据的出现是跨学科技术与应用发展的结果。对于大数据，自然科学家强调网络虚拟环境下对于密集型数据的研究方法，社会科学家则看重密集型数据后面隐藏的价值与推动社会发展的模式。

1.2 研究目标及意义

本白皮书力图从应用、技术、产业、标准等角度，勾画出大数据发展的整体轮廓，探索从应用、技术、产业等维度综合分析大数据标准化工作需求。

本白皮书立足于大数据发展所处的工业社会向信息社会转型的历史时期所独具的政策、经济与文化等特点，分析处于初期发展阶段的大数据对于经济、社会、产业的作用和影响；介绍目前国内、国外主要国家在大数据发展战略、技术与应用方面的布局与实践。

---

[1] Mayer-Schonberger V, Cukier K N. Big Data: A Revolution That Will Transform How We Live, Work, and Think[M]. Eamon Dolan/Houghton Mifflin Harcourt, 2013

本白皮书从数据生存周期的角度提出大数据的技术参考模型，分析大数据发展的关键技术，同时抛开其他影响因素，从数据自身的角度提出在不断创新的应用与服务模式下，大数据的标准体系框架及急需研制的标准项目。

本白皮书的发布旨在与业界分享我们在大数据领域的研究成果、实践经验，呼吁社会各界共同关注大数据的政策研究、技术投入、标准建设与服务应用，共同推动大数据的发展，提升社会整体决策与服务管理能力。

此外，我们还组织了国内相关企业、学校和研究机构的从业人员针对大数据应用、产业、技术与标准化需求进行了问卷调研。总共调研了 28 家高校及科研单位以及 46 家企业（多数为规模在 100 人以上的中大型企业，以 IT 集成商、软件产品提供商为主，同时也包含了一些传统行业和电子商务企业，调研单位清单见附件）。回收有效问卷二百余份[2]。通过对调研数据的分析，初步形成了对于大数据应用、技术、产业发展以及标准化需求的成果，作为业界共同研究的基础。

1.3 指导单位和参与单位

本白皮书的编写得到工业和信息化部软件服务业司和国家标准化管理委员会工业二部的指导，并且也得到了业内有关产、学、研、用等单位和专家的大力支持。北京航空航天大学计算机学院、北京大学信息化与信息管理研究中心、北大方正国际集团、京东商城、中国电子软件研究院、华迪技术有限公司、华为公司、中国农业科学研究院农业信息研究所、北京师范大学管理学院、华中科技大学、武汉大学软件工程国家重点实验室、东方通、上海计算机软件中心、中国石油天然气管道总公司、百度、阿里、腾讯公司、浪潮集团、北京华电祥云、中宇万通、微软中国、甲骨文公司、金蝶公司、IBM 公司等派员参与了本白皮书的编写。

[2]有效问卷数量 204 例。

2.大数据基本概念、特征与作用

2.1 大数据的基本概念和内涵

针对大数据，目前存在多种不同的理解和定义。

按照 NIST 发布的研究报告的定义，大数据是用来描述在我们网络的、数字的、遍布传感器的、信息驱动的世界中呈现出的数据泛滥的常用词语。大量数据资源为解决以前不可能解决的问题带来了可能性。

按照 Gartner 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[3]。

根据百度百科词条的定义，大数据，或称巨量资料，指的是所涉及的资料量规模巨大到无法透过目前主流软件工具，在合理时间内达到撷取、管理、处理、并整理成为帮助企业经营决策更积极目的的资讯[4]。数据规模超出传统数据库软件采集、存储、管理和分析等能力的范畴，多种数据源，多种数据种类和格式冲破传统的结构化数据范畴，社会向着数据驱动型的预测、发展和决策方向转变，决策、组织、业务等行为日益基于数据和客观分析做出。

除了学术界、科研界的定义外，我国 IT 学术界和企业对大数据也有自己的看法，通过调研，我们发现超过一半的受访者认同“新型的数据和分析”，而“新形势的数据应用”和“更大范围的信息”则分列二、三位，“大量的数据”这一选项仅仅位列第四。由此可见，大量的受访者已经意识到大数据的重点在于“数据”的分析和应用，而“大”不过是信息技术不断发展所产生的海量数据的表象而已。(参见图 1)。

[3]引自 Gartner 大数据定义

[4]引自百度百科大数据词条

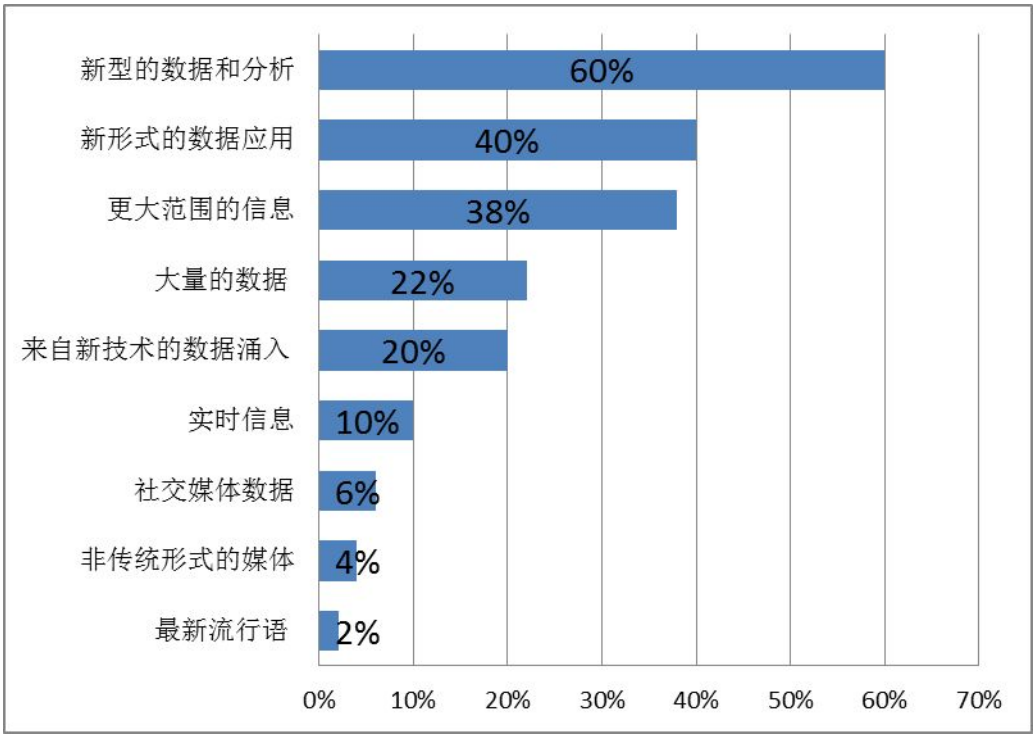

图 1 受访者对于大数据的认识

本报告的观点是大数据代表着数据从量到质的变化过程；代表着数据作为一种资源在经济与社会实践中扮演越来越重要的角色，相关的技术、产业、应用、政策等环境会与之互相影响、互为促进。从技术角度来看，这种数据规模质变后带来新的问题，即数据从静态变为动态，从简单的多维度变成巨量维度，而且其种类日益丰富，超出当前技术与工具控制管理的范畴。这些数据的采集、分析、处理、存储、展现都涉及复杂的多模态高维计算过程，涉及异构媒体的统一语义描述、数据模型、大容量存储建设，涉及多维度数据的特征关联与模拟展现。然而，大数据发展的最终目标还是挖掘其应用价值，没有价值或者没有发现其价值的大数据从某种意义上讲是一种冗余和负担。

2.2 大数据的特征

目前，业内对于大数据特征的研究主要集中在“3V”、“4V”两种，归纳起来，可以分为规模、变化频度、种类和价值密度等几个维度。研究机构 IDC 定义了大数据的四大特征——海量的数据规模、快速的数据流转和动态的数据体系、多样的数据类型和巨大的数据价值，将“价值”作为第四个“V”。其他一些机

构则将真实性作为第四个“V”。还有学者认为应该将（供应商，vendor）作为第四个“V”。

本报告对于大数据的特征从数量（Volume）、多样性（Variety）、速度（Velocity）、价值（Value）以及真实性（Veracity)五个方面进行认识和理解。

数量：聚合在一起供分析的数据规模非常庞大。谷歌执行董事长艾瑞特•施密特曾说，现在全球每两天创造的数据规模等同于从人类文明至2003年间产生的数据量总和。“大”是相对而言的概念，对于搜索引擎，EB属于比较大的规模，但是对于各类数据库或数据分析软件而言，其规模量级会有比较大的差别。

多样性：数据形态多样，从生成类型上分为交易数据、交互数据、传感数据；从数据来源上分为社交媒体、传感器数据、系统数据；从数据格式上分为文本、图片、音频、视频、光谱等；从数据关系上分为结构化、半结构化、非结构化数据；从数据所有者分为公司数据、政府数据、社会数据等。

速度：一方面是数据的增长速度快，另一方面是要求数据访问、处理、交付等速度快。美国的马丁•希尔伯特说，数字数据储量每3年就会翻1倍。人类存储信息的速度比世界经济的增长速度快4倍。

价值：尽管我们拥有大量数据，但是发挥价值的仅是其中非常小的部分。大数据背后潜藏的价值巨大。美国社交网站Facebook有10亿用户，网站对这些用户信息进行分析后，广告商可根据结果精准投放广告。对广告商而言， 10亿用户的数据价值上千亿美元。据资料报道，2012年，运用大数据的世界贸易额已达60亿美元。

真实性：一方面，对于虚拟网络环境下如此大量的数据需要采取措施确保其真实性、客观性，这是大数据技术与业务发展的迫切需求；另一方面，通过大数据分析，真实地还原和预测事物的本来面目也是大数据未来发展的趋势。

在调查过程中，受访者对于大数据特性的关注度如图2所示，从高到低依次为多样性、价值、真实性、数量、速度。

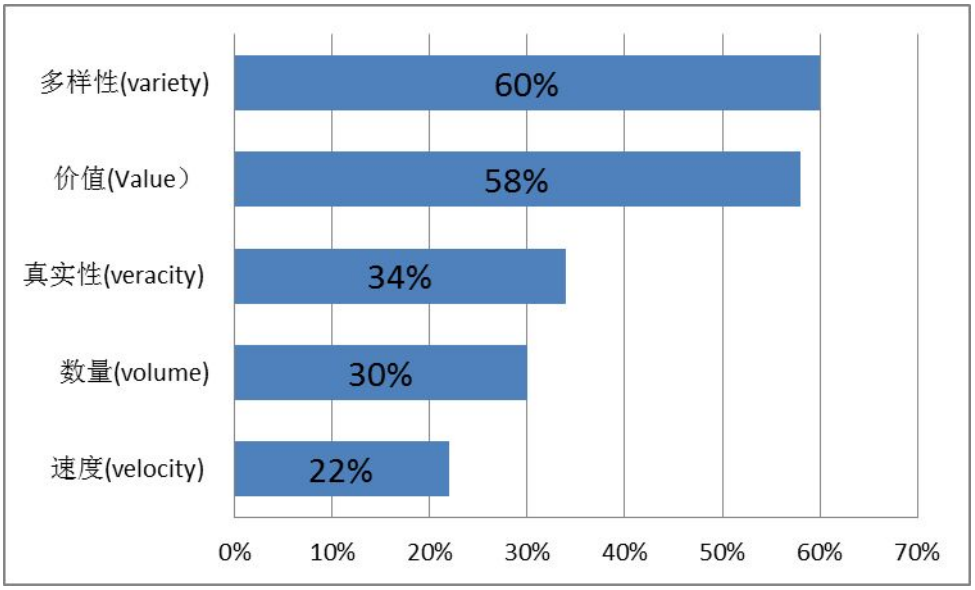

图 2 受访者对于大数据特征的关注度

从图 2 中我们不难看出，在大数据的几个特征中，“多样性”和“价值”最被大家所关注。“多样性”之所以被最为关注，在于数据的多样性使得其存储、应用等各个方面都发生了变化，针对于多样化数据的处理需求也成为了技术重点攻关方向。而“价值”则不言而喻，不论是数据本身的价值还是其中蕴含的价值都是企业、部门、政府机关所希望的。因此，如何将如此多样化的数据转化为有价值的存在，是大数据所要解决的重要问题。

2.3 大数据的重要作用

据资料显示，近年来，甲骨文、IBM、微软、SAP、惠普等公司已经在数据管理和分析领域投入超出 150 亿美元。据 Gartner 最新预测，大数据产业 2014 年将在全球范围内带来近千亿美元的 IT 开支；2015 年，大数据将为全球带来 440 万个 IT 岗位。

1）改变经济社会管理方式

大数据作为一种重要的战略资产，已经不同程度地渗透到每个行业领域和部门，其深度应用不仅有助于企业经营活动，还有利于推动国民经济发展。大数据使经济决策部门可以更敏锐地把握经济走向，制定并实施科学的经济政策。大数

据可以提高企业经营决策水平和效率，推动创新，给企业、行业领域带来价值。大数据技术作为一种重要的信息技术，对于提高安全保障能力、应急能力、优化公共事业服务，提高社会管理水平的作用正在日益凸显。增强安全保障能力。在国防、反恐、安全等领域应用大数据技术，能够对来自于多种渠道的信息快速进行自动分类、整理、分析和反馈，有效解决情报、监视和侦察系统不足等问题，提高国家安全保障能力。

2）促进行业融合发展

网络环境、移动终端随影而行，网上购物、社交网站、电子邮件、微信不可或缺，社会主体的日常活动在虚拟的环境下得到承载和体现。正如工业化时代商品和交易的快速流通催生大规模制造业发展，信息的大量、快速流通将伴随着行业的融合发展，经济形态的大范围变化。虚拟环境下，遵循类似摩尔定律原则增长的海量数据，在技术和业务的促进下，跨领域、跨系统、跨地域的数据共享成为可能，大数据支持着机构业务决策和管理决策的精准性与科学性，社会整体层面的业务协同效率提高。

3）推动产业转型升级

基于传统架构的信息系统很难应付爆发式增长的海量数据，同时传统的商业智能、搜索引擎、分析软件，在面对时空多维度、快速变化的海量数据时，也缺少有效地分析工具、方法和产品。大数据环境下，ICT 产业面临着有效存储、实时分析、高性能计算等挑战，这将对软件产业、芯片以及存储产业产生重要影响。

信息消费作为一种以信息产品和服务为消费对象的活动，覆盖多种服务形态，多种信息产品，多种服务模式。当围绕数据的业务在数据规模、类型和变化速度达到一定程度时，大数据对于产业发展的影响随之显现。

同时，大数据将促进网络通信技术与传统产业更为密切的融合，对于传统产业的转型发展，创造更多价值影响重大。未来，大数据发展将不仅催生软硬件及服务等市场产生大量价值，也将对有关的传统行业转型升级产生重要影响。

4）助力智慧城市建设

信息资源开发利用水平，在某种程度上讲代表着信息时代下社会的整体发展水平和运转效率。大数据与智慧城市是信息化建设的内容与平台，两者互为推动力量。智慧城市是大数据的源头，大数据是智慧城市的内核。仅以智慧交通为例，智慧交通领域的海量数据融合了各类数据，并以城市交通为主题，在海量变化数据中建立关联关系，找到所需数据的准确信息，并被及时推送到对象手中，提高了城市管理的精确性，提升了城市居民的幸福感受。

3.大数据发展现状

3.1 国外大数据发展

大数据发展来源于自然科学、社会科学的技术创新；信息公开、隐私保护、规范管理等的制度建设；各个应用领域主题下的技术路线、模型建设与工具开发等具体实施方案。为此，国外发达国家纷纷提出了大数据的规划、计划、政策以及项目，推动大数据为其国民经济和社会发展服务。

据 IDC 调查分析，目前作为成熟的大数据应用主要集中于欺诈监测、风险管理与商业智能等领域，细分到对于产业，处理与活动领域的大数据应用如图 3 所示。

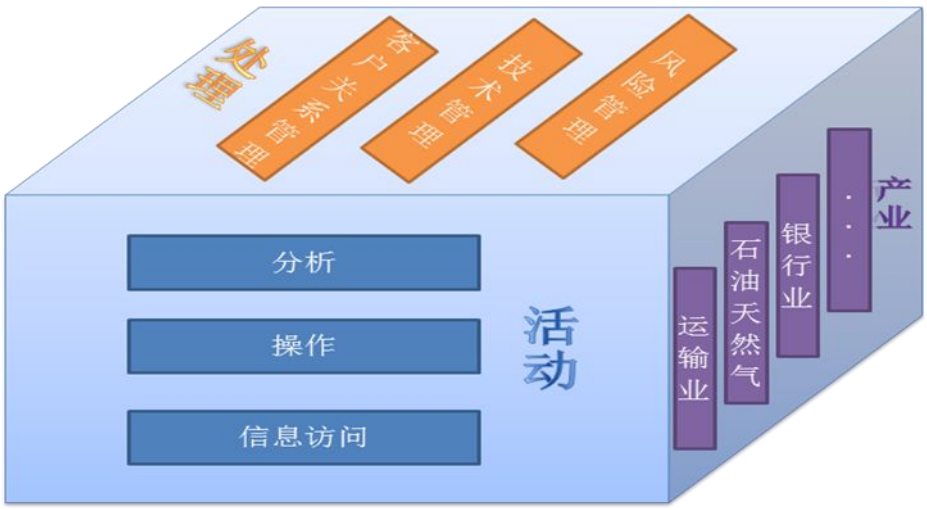

图 3 大数据技术和服务简单用例

图 3 从活动、处理以及产业等三个维度对于大数据技术和服务的相关用例进行了分类。其中活动维度中包括分析（例如数据挖掘、多维分析、数据可视化）、操作（例如运行一个网站、处理网络订单）、信息访问（例如基于搜索的信息获取、规范化,以及内容和数据源的访问）；处理维度包括客户关系管理、供应链和运营、政府、研发、信息技术管理和风险管理；产业维度包括运输行业中的物流优化、零售行业中的价格优化、媒体和娱乐行业中的知识产权管理、石油和天

然气行业中的自然资源勘探、制造业中的保修管理、执法中的犯罪预防和调查、银行业中的欺诈检测、医疗保健行业中的病人治疗和欺诈检测。

对于大数据应用的价值链主要包括三个方面，如图 4 所示：

- 采集与收集：再生老的数据；采集新的数据；提升数据质量。
- 聚合与整合：实时与批量数据的聚合或整合， 多媒体、跨模态数据的聚合；分发给具有弹性计算功能的 IT。
- 消费与应用：商业智能 BI 或数据仓库 DW 的集成；可视化；业务集成。

图 4 数据和风险管理中的大数据价值链

3.1.1 政府出台计划

3.1.1.1 美国

2011 年，“总统科学技术顾问委员会（President’s Council of Advisorson Science and Technology）建议”认为大数据相关技术具有重要战略价值，而联邦政府对其研发投资不足。作为建议的反馈，白宫科技政策办公室发布了《大数据研究和发展倡议》，并组织了大数据高级监督小组（Senior Steering Group on Big Data）协调和拓展政府在这一重要领域的提升美国利用收集的庞大而复杂的数字资料提炼真知灼见的能力，协助加速科学、工程领域的创新步伐，强化美国国土安全，转变教育和学习模式。

《大数据研究和发展倡议》提出联邦政府希望与行业、科研院校和非盈利机构一起，共同迎接大数据所创造的机遇和挑战。某种程度上，大数据技术在美国已经形成了全体动员的格局，并承诺将在科学研究、环境保护、生物医药研究、教育以及国家安全等领域利用大数据技术进行突破。同时，美国国家科学基金会（NSF）、国家卫生研究院（NIH）、国防部（DOD）、能源部（DOE）、国防部高级研究局（DARPA）、地质勘探局（USGS）等六个联邦部门和机构承诺，将投入超过 2 亿美元资金用于研发“从海量数据信息中获取知识所必需的工具和技能”，并披露了多项正在进行中的联邦政府计划，主要内容如下：美国国家科学基金和美国国家卫生研究院主要推进大数据科学和工程的核心方法及技术研究，项目包括管理、分析、可视化、以及从大量的多样化数据集中提取有用信息的核心科学技术；国防部高级研究局项目主要推进大数据辅助决策，集中在情报、侦查、网络间谍等方面，汇集传感器、感知能力和决策支持建立真正的自治系统，实现操作和决策的自动化；美国能源部试图通过先进的计算进行科学发现，提供 2500 万美元基金来建立可扩展的数据管理、分析和可视化研究所。美国地质勘探局通过给科学家提供深入分析的场所和时间、最高水平的计算能力和理解大数据集的协作工具，催化在地理系统科学的创新思维。

2012 年 3 月，美国白宫宣布启动大数据研究和开发，6 月，美国国家标准技术研究所（NIST）启动了大数据相关研究。2013 年 6 月，NIST 召开了大数据公共工作组（Big Data Public Working Group ，BD-PWG）成立会议，并于 9 月启动了大数据定义和数据、通用需求、参考架构、安全隐私及技术路线图等内容的研究，并提出了《大数据参考架构》报告，受到多方面关注。

3.1.1.2 欧盟

2010 年 11 月，欧盟委员会提出了“欧盟开放数据战略”，旨在将公共部门搜集和产生的原始数据通过再利用成为数以万计 ICT 用户依赖的数据材料，同年 12 月正式推进这一战略并提出有关开放数据战略的多项法律提案，提案指出：“所有来自于公共部门的文件除非受第三方版权保护外均可用于任何目的（商业

或非商业），大部分公共部门的数据都将免费或几乎免费，强制要求提供通用的且机器可读格式的数据，确保数据的有效再利用，数据开放范围将覆盖包括图书馆、博物馆、档案馆等在内的更广泛的组织”。

“欧盟开放数据战略”将重点加强在数据处理技术、数据门户网站和科研数据基础设施三方面的投入，旨在欧洲企业与市民能自由获取欧盟公共管理部门的所有信息，建立一个汇集不同成员国以及欧洲机构数据的“泛欧门户”。

未来，欧盟开放数据战略将重点加强在数据处理技术、数据门户网站和科研数据基础设施三方面的投入。目前比较成功的应用有“你的议会”（www.itsyourparliament.eu），公民可以通过该网站了解欧洲议会的选票情况，查看投票记录并投票；英国制药（www.data.gov.uk/apps/uk-pharmacy），通过智能手机帮助市民在英国找到距离最近的药店；欧洲能源（http://energy.publicdata.eu/ee/vis.html），对欧盟统计局和其他机构提供的数据进行加工，将欧洲能源消费情况可视化；开放企业（http://www.opencorporates.com），是关于公司的数据库，目前已包含超过 30 个地区 3000 万家企业的 URL。

3.1.1.3 联合国

联合国推出了名为“全球脉动”（Global Pulse）的新项目，希望利用“大数据”预测某些地区的失业率、支出削减或是疾病爆发等现象。

全球脉动技术的目标在于利用数字化的早期预警分析，来提前规划、调整、指导联合国在全球范围内，针对众多行业领域的援助项目，以提高援助项目完成的精确性和有效性。

3.1.1.4 多国联盟

合作下的数据开放是目前的潮流，也是大数据应用的前提。2011 年美国、英国、巴西、挪威、墨西哥、印尼、菲律宾、南非等八国宣布成立“开放政府联盟”（OGP），并发布《开放政府宣言》，宣言书说：“政府代表公民收集并保

存各种各样的信息，公民有权利获取关于政府活动的各种信息。我们承诺：用可以重复使用的格式，及时主动地向社会开放高质量的信息，包括原始的数据。”

2011 年 12 月，美国联邦政府宣布将和印度政府共同合作，把现有的 Data.gov 改造成开源平台，印度将率先移植 Data.gov，作为其中央政府的数据开放平台。

英国政府自 2011 年 11 月发布了对公开数据进行研究的战略政策，同时致力于探索公开数据在商业创新和刺激经济增长方面的潜力。

英国政府投资支持成立开放式数据研究所 ODI(The Open Data Institute)。未来，英国政府将通过这个组织来利用和挖掘公开数据的商业潜力，并为英国公共部门、学术机构等方面的创新发展提供“孵化环境”，同时为国家可持续发展政策提供帮助。

法国政府在《数字化路线图》中列出了五项将会大力支持的战略性高新技术，而“大数据”是其重要内容。2013 年 4 月法国政府召开“第二届巴黎大数据大会，”，会上法国经济、财政和工业部门宣布将投入 1150 万欧元用于支持 7 个未来重点项目。这些项目的目的在于“通过发展创新性解决方案，并将其用于实践，来促进法国在大数据领域的发展。”

此前，法国软件编辑联盟(AFDEL)曾号召政府部门和私人企业共同合作，投入 3 亿欧元用于推动大数据领域的发展。AFDEL 认为，未来 5 年内，大数据创造的价值将会达到 28 亿欧元，同时将会产生 1 万个工作岗位。

3.1.2 工业界大数据研究

工业界针对大数据分析平台，纷纷推出自己的大数据分析工具，主流的平台和产品如下：

3.1.2.1 Google 的大数据分析产品

Google 公司作为全球最大的信息检索公司，走在了大数据研究的前沿。面对呈现爆炸式增加的因特网信息，仅仅依靠提高服务器性能已经远远不能满足业

务的需求。如果将各种大数据应用比作“汽车”，支撑起这些“汽车”运行的“高速公路”就是云计算。正是云计算技术在数据存储、管理与分析等方面的支持，才使得大数据有用武之地。Google 公司从横向进行扩展，通过采用廉价的计算机节点集群，改写软件，使之能够在集群上并行执行，解决海量数据的存储和检索功能。Google 公司大数据处理的几大关键技术为：Google 文件系统 GFS、MapReduce、Bigtable 和 BigQuery。Google 的技术方案为其他公司提供了一个很好的参考方案，各大公司纷纷提出了自己的大数据处理平台，采用的技术也都大同小异。

3.1.2.2 惠普的 HAVEn

HAVEn 平台提供了大量的应用开发接口（API），惠普希望通过 HAVEn 与合作伙伴共同打造一套完整的大数据分析生态系统，让更多应用解决方案落地到行业。它可以充分利用惠普的分析软件、硬件和服务，创建新一代为大数据准备的分析应用和解决方案。“HAVEn”这个名字实际上来源于其各个组件的首字母，即 Hadoop（HDFS）、Autonomy、Vertica、Enterprise Security 以及 nApp（行业解决方案）。可以看出，HAVEn 平台实际上是一个惠普大数据产品的组合。具体而言，HAVEn 并不是简单的产品堆叠，惠普对其中各个组件的交互与连接都进行了设计与优化，并提供了统一的框架。HAVEn 平台能够从各种数据源进行集成，分析各种类型数据，如传统数据仓库、机器生产数据、电子邮件、文本数据以及企业外部的社交媒体数据。

3.1.2.3 Teradata

日前，全球领先的大数据分析和数据仓库解决方案厂商 Teradata 天睿公司发布了 Teradata Aster 大数据综合分析平台。作为业内首款整合大数据分析平台，实现了将开源 Apache Hadoop 和 Teradata Aster 整合至高度集成和优化的单一平台中。该平台采用 Teradata Aster 的 SQL-MapReduce 和 Aster SQL-H 专利技术，支持用户透明地访问 Hadoop 平台，为广大知识型员工提供独特的业务分析功能。该平台预先封装多项即开即用的分析功能，能够在数小时内快速实现

数字营销优化、社交网络分析、欺诈侦测以及机器生成数据的分析等。Teradata Aster 大数据综合分析平台专为满足苛刻的分析需求设计，提供更强的计算能力、更大的内存容量及更快的数据移动。同市场上其他典型平台相比，该平台的数据吞吐量及分析速度分别提高 19 倍及 35 倍。Teradata Aster 大数据综合分析平台配备充足的内存和高速宽带互联功能，能够支持极度密集的复杂分析计算，相比现有其他产品更加简洁。采用 Teradata Aster 大数据综合分析平台后，用户无需复杂的培训即可使用 MapReduce 和 Hadoop 技术。

3.1.2.4 IBM 的 InfoSphere

2011 年 5 月，IBM 正式推出 InfoSphere 大数据分析平台。InfoSphere 大数据分析平台包括 BigInsights 和 Streams，二者互补，Biglnsights 对大规模的静态数据进行分析，它提供多节点的分布式计算，可以随时增加节点，提升数据处理能力。Streams 采用内存计算方式分析实时数据。InfoSphere 大数据分析平台还集成了数据仓库、数据库、数据集成、业务流程管理等组件。BigInsights 基于 Hadoop，增加了文本分析、统计决策工具，同时在可靠性、安全性、易用性、管理性方面提供了工具，并且可与 DB2、Netezza 等集成，这使大数据平台更适合企业级的应用。比如，BigInsights 提供了一种类似 SQL 的更高级的查询语言。再如，除了支持 Hadoop 的 HDFS 存储系统外，BigInsights 还支持 IBM 最新推出的 GPFS SNC 平台，以更好地利用其强大的灾难恢复、高可靠性、高扩展性的优势。企业级产品更重要的是没有单点故障，GPFS 让整个分布式系统更可靠。Hadoop 本身不提供分析的功能，因此 BigInsights 平台增加了文本分析、统计分析工具。

3.2 国内大数据现状

国家在推进信息化、电子政务、智慧城市等领域发展与建设，多次强调要重视整体提升信息资源开发利用水平，强调要关注并重视大数据工作。

目前，国内对于大数据的实质推进更多地处于科研、应用、地方、产业等部门单个探索实践中。部分信息化发展基础较好的地方，其信息化发展规划及产业部署中已经明确将推动大数据的发展与应用。

3.2.1 国内大数据关注焦点

通过调研显示，目前在大数据的行业领域应用关注度上，“智慧城市”、“政务”以及“公共服务”位列前三。（见图 5）

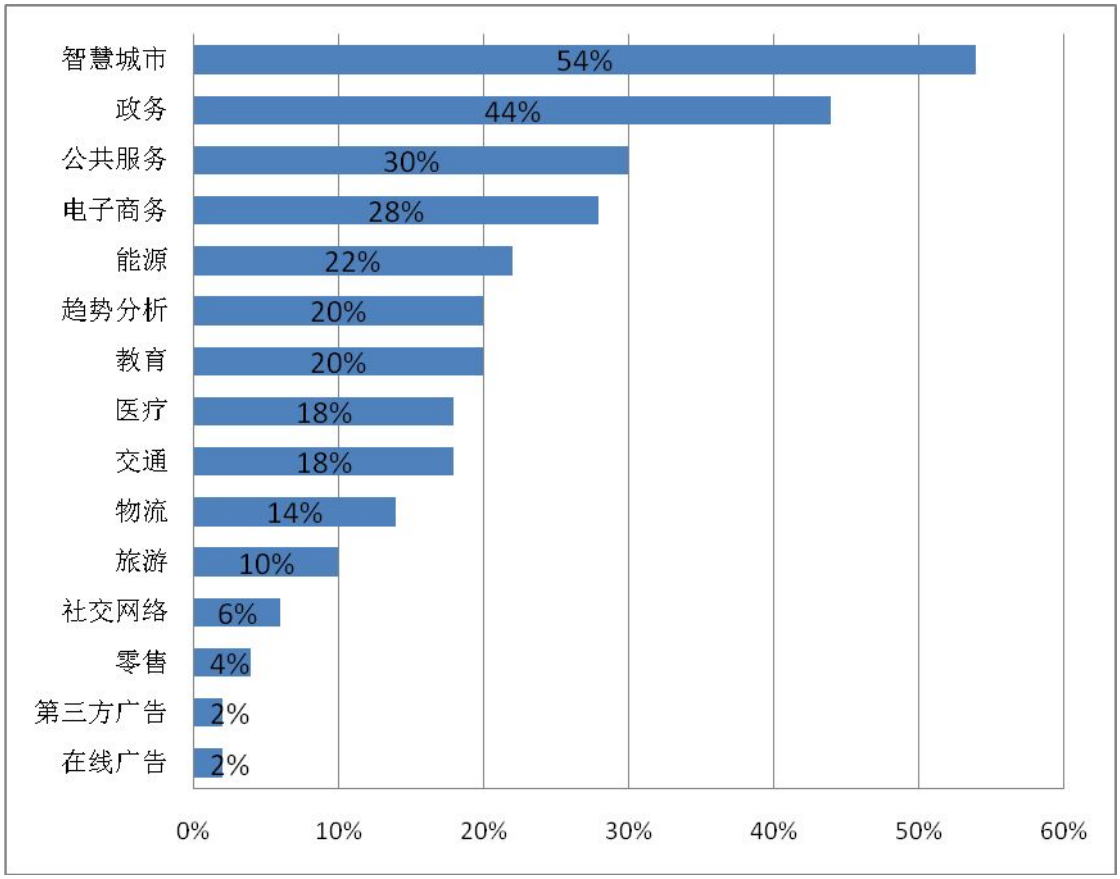

图 5 各领域的大数据关注度

不难看出，目前对于大数据应用有迫切需求的主要集中在政府部门。政府部门在推动社会管理与公共服务过程中，希望通过对于现有的和正在产生的大量、多媒体的数据进行有效的分析和应用，支持基础设施建设和提高服务水平。对于“能源”、“教育”、“医疗”、“交通”等领域的大数据关注度大体相当，体现了大数据应用的广阔性；这些领域在传统业务推进中头绪比较复杂，数据资源

开发水平低，科学化决策难度大，大数据的发展应用在某种程度上增强了对于复杂形势的分析，加强了对于科学决策的客观数据支持，在这些领域中大数据应用前景广阔。

在具体技术层面，“信息集成”成为了国内大数据关注的重点。目前大部分单位及受访者都表示已经利用一个集成的、可缩放的、可扩展的和安全的信息基础设施开始推动大数据应用实践。同时，在实践过程中对于数据的安全性与治理、大容量的数据存储与管理、基础架构、相关工具等也是大数据关注的重要技术领域。（见图 6）

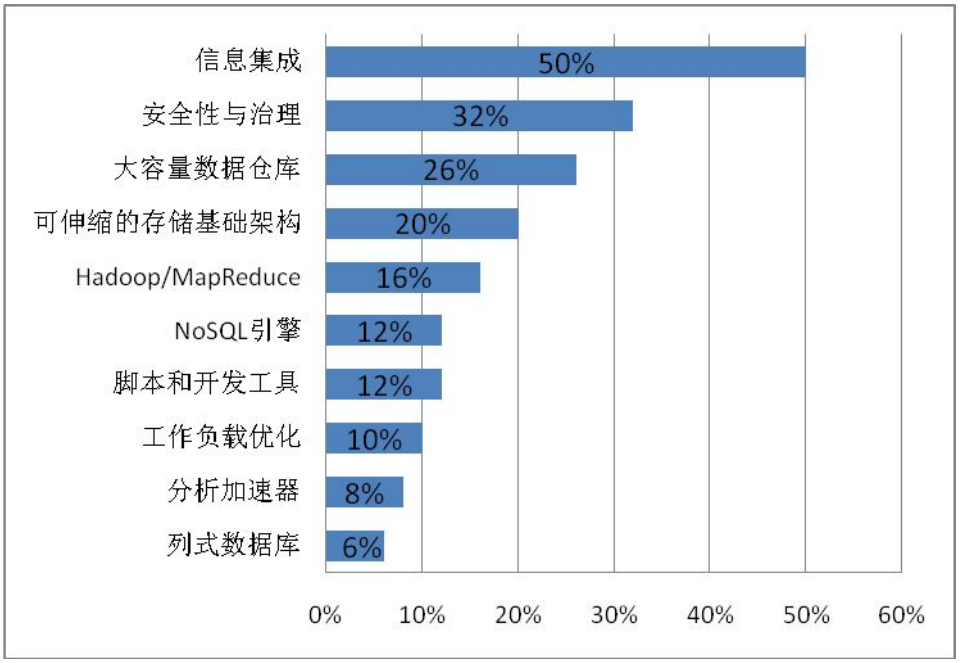

图 6 大数据技术关注度

3.2.2 地方政府和组织探索大数据应用

3.2.2.1 北京市

北京市经信委牵头，北京市各政务部门共同参与建设了北京市政务数据资源网（www.bjdata.gov.cn），于 2012 年 10 月推出测试版，目前正在试运行基础上加快制定管理办法。

目前，北京市已有 29 个部门公布了 400 余个数据包，涵盖旅游、教育、交通、医疗等各个门类。打开网站主页可以看到，点击量最高的是“土地用途分区”，已被下载 435 次，由北京市国土资源局提供。旅行社、机场班车线路、星级饭店、高校信息也是非常热门的下载资源。北京市政务数据资源网正在面向企业及个人征集 APP（应用程序），一些社会力量开发的 APP 正在进行技术测试和审查。

在该网站可以看到，“游北京”和“爱健康”目前已经可以下载试用。前者可以查阅北京旅游景点、餐饮、促销信息、洗手间信息等，后者是北京市所有卫生保健设施的指南应用，包括诊所、医院、养老院等信息，用户可以利用这款软件定位附近的医疗设施，查看现场网络图像。

3.2.2.2 上海市

上海目前正在加强对大数据领域的深化研究。2013 年，上海启动推进大数据研究与发展的三年行动计划，重点选取医疗卫生、食品安全、终身教育、智慧交通、公共安全、科技服务等具有大数据基础的领域探索建设大数据公共服务平台。上海市政府于 2014 年 5 月明确上海将率先实行政府数据资源向社会开放，出自 28 个市级政府部门、涵盖 11 个领域的 190 项数据内容将成为今年重点开放对象——从医院床位信息到候诊人数信息，从挖路、占路、封路信息到停车场库及路侧车位信息，政府大数据“富矿”可供全民开采。国内首个政府数据服务网 www.datashanghai.gov.cn 作为开放统一入口，提供数据查询、浏览、下载等功能。而且，上海将重点建设政府移动 APP 门户，将各部门开发的各类公众服务 APP“一网打尽”，让市民通过这个门户方便地检索和下载所需 APP。

此前，上海已启动政府数据资源向社会开放试点，建成“上海政府数据服务网（一期）”，9 家试点单位开放的数据产品及应用，涵盖地理位置、道路交通、公共服务、经济统计、资格资质、行政管理等 6 大领域。如市商务委开放了内贸、外贸、外资、外经及综合 5 类数据产品；市交通委提供了全市搬场企业名录、全市公交枢纽站分布、中心城区公交站点分布、停车场位置分布等；市住房保障管理局开放了保障房工程信息、房地产开发企业信息、房地产经纪企业信息等。在

此基础上，上海市经信委通过印发《2014 年度上海市政府数据资源向社会开放工作计划》，规定再开放公共安全、公共服务、交通服务、教育科技、金融服务、能源环境、健康卫生、文化娱乐等 11 个领域，开放的市级政府部门数量是原先的 3 倍。其中，地理位置类的数据资源将全面开放，市场监管类数据也成为开放重点，并大力推进交通数据资源开放。根据市政府总体规划，政府数据资源开放主体未来将扩展到法律法规授权的具有管理公共事务职能的组织，以及与人民群众利益密切相关的公共企事业单位。按照年度计划，上海将参照图书资源的管理模式，力争 3 年内，完成各政府部门信息系统所承载的信息资源分类、目录编制注册，实现全市政府数据资源目录的集中存储和统一管理，基本摸清政府数据资源家底。

同时，上海市经信委正在研究成立大数据局，成立后将推进上海政府层面的数据公开和信息共享，以解决政府信息资源家底不清、认识不够以及部门间数据信息共享不充分等目前上海在数据资源管理和运用上存在的问题。

3.2.2.3 广东省

广东省是国内率先关注并推动大数据的地方之一。2013 年 5 月出台《广东省信息化发展规划纲要（2013-2020 年）》，在智慧广东建设任务中，该纲要明确“到 2015 年，全省信息化总体达到中等发达国家水平，珠三角地区信息化水平迈进世界先进行列。智慧城市建设取得显著成效，信息基础设施进一步完善，信息技术自主创新体系基本形成，信息技术与传统产业深度融合，大数据和商业智能试点示范应用成效明显，公共服务和社会管理电子化、网络化全面普及，信息化有效推动产业转型升级和生产方式转变，信息化成果惠及全省人民。”在构建信息技术产业体系的发展任务中，该纲要明确“构建面向企业经营管理及社会服务和管理的大数据挖掘应用创新平台，并以广州、深圳两大超级计算中心为基础构建信息技术研发设计、高性能计算创新平台。”在推动信息化和工业化深度融合发展任务中，该纲要明确“推进大数据商业化应用。充分利用市场机制，加快推进行业、企业开展大数据应用。支持和鼓励行业协会、中介组织开发深度加工的行业应用数据库，建立行业应用和商业服务大数据公共服务平台，提供数据

挖掘分析和商业智能等大数据应用服务，帮助中小微企业定制各类大数据应用解决方案。培育数据资源服务重点企业，提高数据资源服务能力。推动大数据在生产过程中的应用，鼓励企业运用大数据开展个性化制造，创新生产管理模式，降低生产成本，提高企业竞争力。加快商业大数据创新应用，鼓励企业开展精准营销、个性化服务，提高流通、销售等环节的管理水平。”在推进城镇管理和服务智慧化任务中，该纲要明确提出“建设智慧城镇运营平台，建立健全数据采集、交换共享、开发利用相关标准体系，开展智慧城镇大数据应用，推动城镇创新发展。深入推进智慧城市试点建设，引导全省智慧城市建设有序推进。”

早在 2012 年广东省经济和信息化委员会就开展了“广东省实施大数据战略工作方案”的研究，立足于坚持以开放共享推动大数据应用，以开发应用带动大数据发展，以大数据发展促进社会创新，建成智慧广东。方案中提出，为保证大数据战略有效实施，将建设政务数据中心，并为高等院校和企业等成立大数据研究机构提供支持；将在政府各部门开展数据开放试点，并通过部门网站向社会开放可供下载和分析使用的数据，进一步推进政务公开。

3.2.2.4 陕西省

近年来，陕西省电子政务与信息化建设快速推进。一方面，加强了顶层设计和集中部署，另一方面电子政务公共平台服务体系初步建成。陕西省各级政府及相关部门的信息化服务，不再需要重复建设网络、机房，不再考虑存储、灾备等因素。

2012 年 12 月陕西省发布了“大数据产业发展战略”与“沣西大数据产业园发展规划”。陕西省大数据产业发展分为三个阶段：2012-2013 年是导入期，以建设政务公共平台为支撑，以政务信息资源建设服务为基础，构建基于高性能计算的大数据计算处理平台和环境；2013-2015 年是建设期，也是战略机遇期，根据人口、林业、社情民意调查分析、社会管理与服务、金融等领域对大数据处理需求，承接其他国家有关部委和央企数据中心或灾备中心落户，形成大数据产业洼地，将全国人口信息处理与备份中心落户西咸新区作为陕西发展大数据产业的

重要机遇；2016-2017 年是成长期，围绕国家基础数据的上下游流入，形成以政务大数据服务产业为核心的高黏性信息服务产业生态。到 2017 年，建成以西咸新区为核心的国家级大数据处理与服务产业集群，成为国家政务信息资源的汇集地、社会信息资源的集散地。

沣西大数据产业园选址位于西咸新区信息产业园内，总占地约 5 平方公里，拟分三大板块推动大数据产业发展。第一板块为数据基础层产业集聚区，第二板块为软件开发和信息服务集聚区，第三板块为预留拓展区，作为未来信息产业持续增长的重要保障。目前中国移动、中国电信、中国联通三大运营商以及全国人口数据处理与备份(西安)中心项目已经入区，产业集聚初具规模。“沣西大数据产业园发展规划”以“数据沣西、智慧西咸、备份中国、物联世界”为目标，以实现数据的“规模化集中吞吐、深层次整合分析、多领域社会应用、高效益持续增值”为方向，大力发展数据存储、呼叫中心、IDC 中心、灾备中心、数据交换共享平台等业态，积极创新商业模式。园区建设划分为三个阶段:

第一阶段(2013—2015 年)为培育期，即基础网络和数据中心建设期。本阶段主要依托四大运营商数据中心、基础网络层的建设，构建海量存储和高速传输网络，为大数据产业发展提供基础和保障。同时，全力引进人口信息、自然资源和空间地理信息、法人单位数据、宏观经济数据等国家基础、专有的数据资源存储与服务中心，数据灾备基地和国家超级计算中心落户，在业内举起大数据处理与信息服务产业的旗帜，到 2015 年底完成固定资产投资 100 亿元，初步建成大数据产业发展的网络传输平台和基础信息资源集聚区。

第二阶段(2016—2017 年)为成长期，引进龙头企业，培育数据分析企业。制定政策机制，完善园区规划，构建宽松发展环境，全力引进微软、IBM、惠普、谷歌、百度、阿里巴巴、腾讯、世纪互联等数据存储、分析和应用企业入园，集聚一批国内外龙头企业的研发总部、技术中心和高端制造部分，重点培育一批数据分析企业。到 2017 年底累计完成固定资产投资 200 亿元、实现产值 200 亿元，形成相对完整的数据服务产业集群，建成“陕西大数据处理与服务产业基地”，

力争率先建成数据应用示范区，推动园区进入国家级信息产业基地行列，实现产业和城市的优化升级。

第三阶段(2018—2020 年)为成熟期，整合数据资源，形成以数据为基础的信息服务产业特色。依托“云计算”和“物联网”等着力点，进一步整合多领域数据资源，探索以数据资源为基础的信息服务产业发展模式，研究和规范数据资源的应用、范围和权限等，为信息服务产业大发展奠定基础，引领社会化信息服务模式的全面提升。到 2020 年底实现产值 500 亿元，聚集各类人才 5 万以上，使沣西新城成为国家级的信息产业园和大数据处理中心。

图 7 沣西大数据产业园发展规划图

3.2.2.5 贵州省

贵州省也在积极布局大数据产业。从 2013 年开始，贵州发力大数据产业布局。中国电信、联通、移动三大电信运营商数据中心在贵州开工建设、中关村贵阳科技园成立、富士康第四代产业园落户等等一系列大手笔，正助推贵州迈上“云端”，成为发展大数据产业的黄金宝地。

从 2013 年下半年开始，三大电信运营商在贵安新区分别建设的全国性数据中心，计划总投资 100 多亿元，总规模将超过 10 万个机柜，服务器将超过 200 万台。数据中心建成后，将形成超过 2500PB 的裸容量存储能力，730 亿 TPCC 计算能力，可为大数据提供巨大存储服务和计算服务，将为新区加快大数据及其关联产业发展奠定坚实基础。三大运营商的数据中心在贵安新区相继建成后，将使贵阳周边特定区域快速集聚 20 万—30 万的机架、百万台的服务器，数据存储规模可达 EB 以上，随着大数据产业持续发酵，将形成一个千万服务器集群的数据中心基地，是国内乃至全球最大的数据聚集地之一。

2013 年 3 月 1 日，贵州·北京大数据产业发展推介会在北京举行，大数据产业联盟与贵州省政府签订合作框架协议，共同推动大数据产业的发展。其后两天，中关村贵阳科技园清镇园招商引资推介会在京举行，16 个项目在推介会上成功签约。2013 年 9 月 8 日，“中关村贵阳科技园”揭牌，为贵阳市发展新一代信息技术产业提供重要支撑，也为大数据产业的发展提供强大支撑。

2013 年 10 月，富士康(贵州)第四代绿色产业园一期项目在贵阳市贵安新区开工。

步入 2014 年，贵州在大数据产业持续发力。2014 年 3 月，贵州省颁布了《关于加快大数据产业发展及应用若干政策的意见》和《贵州省大数据产业发展应用规划纲要（2014-2020 年)》，抛出了 30 条鼓励措施，明确从今年起连续 3 年，省和贵阳市、贵安新区每年各安排不少于 1 亿元资金，用于支持大数据产业发展及应用。

2014 年 6 月底，贵州省大数据产业研究院将在贵州大学开工建设的消息在贵州省大数据产业发展应用研究院研讨会上发布。该研究院由贵安新区管委会、泰豪集团有限公司、贵州大学等联合建设，建设总经费 15000 万元，由研发大楼、综合办公楼、产业孵化楼和公寓等组成，实行理事会领导下的院长负责制。

2014 年 7 月 11 日，云上贵州·大数据国际年会论坛举行,论坛围绕“大数据时代的产业变革——融合创新、绿色跨越”的主题,采取“主题演讲、高峰论

坛、媒体活动、参观调研”相结合的形式，邀请国家部级领导、专家院士、企业家代表等，共同探讨大数据产业发展面临的机遇与挑战、趋势与未来，共同研究贵州大数据产业发展的方向和路径选择。同时，正式启动了贵州省大数据产业联盟。

此外，贵阳市政府将与北京云基地共同建设贵阳云基地，联合完成贵阳市云计算产业发展规划，启动贵阳云计算产业创业投资基金、云计算设备生产基地、云计算创新孵化基地、国际云服务数据中心基地、电子政务云示范等项目的建设。到 2016 年聚集 50 家以上云计算企业，形成服务器年产值 50 亿元，云计算应用产值 10 亿元。

3.2.2.6 产业联盟

各地方结合其经济、技术、产业等发展需求，以产业联盟等形式推动大数据发展。

2012 年 12 月 13 日，中关村大数据产业联盟是成立。联盟成立宗旨是：把握云计算、大数据与产业革新浪潮带来的战略机遇，聚合厂商、用户、投资机构、院校与研究机构、政府部门的力量，通过研讨交流、数据共享、联合开发、推广应用、产业标准制定与推行、联合人才培养、业务与投资合作、促进政策支持等工作，推进实现数据开发共享，并形成相关技术与产业的突破性创新，产业的跨越式发展，推动培育世界领先的大数据技术、产品、产业和市场。联盟主要会员单位包括：中国宽带资本、亿赞普（北京）科技有限公司、天云融创数据科技（北京）有限公司、北京友友天宇系统技术有限公司、百度在线网络技术（北京）有限公司、北京东方国信科技有限公司、北京拓尔思信息技术有限公司等。联盟主要工作包括承接政府部门的相关大数据方面的产业课题研究项目、组织相关大数据产业国际会议、大数据专业书籍的编著及出版、组织大数据培训项目开发、组织联盟成员进行政府项目申报、联盟间的合作、大数据项目孵化及标准研究、推动联盟间的国际合作和搭建数据市场服务平台。

2013 年 3 月 28 日深圳市大数据产业研联盟成立。发起成立联盟的 16 个单位主要包括：中科院深圳先进技术研究院、国家超级计算深圳中心（深圳云计算中心）、深圳大学、清华大学深圳研究生院、深圳市南山科技事务所、金蝶国际软件集团有限公司、华为技术有限公司、宇龙计算机通信科技（深圳）有限公司、华大基因、腾讯公司等。该联盟立足于发挥深圳高新技术研究和产业化优势，发挥产业联动作用，促进同行业间信息沟通、业务合作、资源共享、优势互补，促进大数据产业链的形成。

2013 年 6 月，山东农业大数据产业技术创新战略联盟成立。包括 6 个省直厅局，2 所农业高校，2 家科研单位，还有 11 家计算机和信息技术、农业产业方面的国内知名企业作为联盟的成员。这个由政府、高校、科研单位、企业组成的联盟将通过加强对农业相关信息和数据的分析研究，为政府决策、产业发展提供更多的服务和支持。联盟成员包括：山东省科技厅、教育厅、农业厅、林业厅、国土资源厅、水利厅、省畜牧局、农机局等政府部门，中国测绘科学研究院、山东省农科院、山东农业大学、青岛农业大学等科研单位和高校，以及龙信数据（北京）有限公司、浪潮集团、山东金正大生态工程股份有限公司、山东登海种业股份有限公司等国内企业。与农业相关的信息、数据来源十分广泛，包括气象、土地、水利、农资、农业科研成果、动物和植物生产发展情况、农业机械、病虫害防治、生态环境、市场营销、食品安全、公共卫生、农产品加工等诸多环节。这里更为重要的是农业大数据的应用。联盟将致力于成员间的沟通与合作，以联盟为沟通与合作平台，共同围绕农业大数据产业技术创新的关键问题，加强合作，联合攻关；加强联盟创新资源的整合与共享，强化创新人才培养机制创新，积极推进大数据研究的学科建设；探索建立互利互惠、富有生命力、符合市场化和产业化的运行机制；积极开展农业大数据的示范推广，为政府部门科学决策提供参考。

3.2.3 国内大数据应用实践

大数据在国际上已经有了很广泛的应用，并带来了巨大的经济效益。国内大量企业纷纷意识到，随着大数据相关技术的不断发展，传统的商业模式将被颠覆，

新的商业生态将形成，而且随着价值链各方对业务模式和盈利模式的创新，新的商业生态将在不断演化中完善。因此各个企业纷纷开展自己的大数据布局。

目前大数据在国内各行各业也得到了广泛的应用。包括电子地图，电子商务、电信、互联网、媒资、高性能计算、金融等行业和领域都有应用。本报告针对国内大数据在地图数据，电子商务和科学研究领域的典型应用案例进行了分析。

3.2.3.1 地图数据领域

高德地图作为数字地图、导航和位置服务解决方案提供商，掌握了大量的行业运营车辆 GPS 数据，以及高德用户数据，并与各城市交管部门合作，掌握了众多交通信息数据。

高德和阿里巴巴开展了数据领域的共享合作。在数据交换方面，两家公司拟联合建立数据库系统，高德提供地理位置、交通信息、兴趣点信息（point of interest, POI）以及用户数据等，阿里巴巴则分享其电商平台如淘宝、天猫上商家的地理位置信息以及其他基于网络的地理位置信息从而解决两家公司间各自数据来源领域的不同所导致的数据单一等问题。通过进行充分有效的数据共享交换，使得两家公司的数据资源都得到了充分的补充和扩展，为之后进一步的数据挖掘和分析提供了一个良好的环境。两家企业未来将共建大数据服务体系，高德拥有基础的地图和导航数据，阿里巴巴在电子商务尤其是商户信息方面非常强大，其电子商务平台上每天有上千万商户交易、物流配送等信息，未来两家企业会把各自的数据融合、匹配，建立大数据服务体系。两家在数据服务上与其他传统的技术厂商之间将产生巨大的差距，其合作优势将会很快在服务中得到展现。

除与阿里巴巴的合作之外，高德地图还与嘀嘀打车、团 800、大众点评、携程、丁丁优惠、订餐小秘书等第三方资源进行合作。通过与这些第三方资源的数据开放和共享，一方面提高高德地图本身的数据来源和储备，为其服务提供更加有力的支持。同时高德地图也将其自身的数据与这些企业进行共享，从而带动这些企业相关业务的开展。

3.2.3.2 电子商务领域

拥有十年电商经历的京东积累了非常多的有价值的数据，京东已经吸引了一大批非常有质量的用户，上亿用户的数据对于任何一家电商来说，都具有一定的价值。

目前京东将其交易、营销、供应链、仓储、配送、售后和 IT 等七大系统所产生的数据，通过其数据平台全面开放。提供超过 500 个 API 的调用，用户可以通过调用其提供的 API 来获得在京东大数据平台上的相关数据，从而为其相关的应用提供便利。目前京东数据平台的开放 API 的日均调用量超过了 2 亿次，合作的 ISV（Independent Software Vendors 独立软件开发商）500 多家，注册的个人用户达到了 3 万人次。为一万余家商户提供了服务。

京东作为国内知名的电子商务平台，其数据服务的主要对象就是在其平台上进行销售的商户和购买商品的客户。因此京东提供的数据服务，重点集中在以下几个方面。

1）精准营销

几乎所有的电商企业都会基于用户的购买行为做精准营销，主要方式是 E-mail、短信等。网站推介系统也是一种较为隐蔽的营销方式。京东依靠大数据进行精准营销，最重要的是用户建模。

2）优化供应链

京东的很多商品都是自动补货，系统会根据销售情况和市场预期，依靠预测模型，在库存量达到某一个阀值时自动生成订单发给供货商。一些复杂的因素会被去除掉，例如团购等，以保证预测模型的准确。

大数据也被应用在物流配送领域。京东会分析物流人员、仓库以及用户之间的地理关系，为物流人员提供最优配送路径，提高配送速度，提升用户体验。

3）智能网站

基于大数据挖掘和分析，网站将变得越来越智慧。一些商品具有重复购买的特点，例如牙膏，购买之后在可预期的一段时间内将会用完。京东会分析此类商品用户两次购买之间的平均时间，在这个时间到来之后，推介系统有可能会给用户推介相应的商品，提升用户的体验，提高商品的转化率。

3.2.3.3 科学研究领域

中国科学院计算机网络信息中心研发了中科院科学数据库。截止到 2010 年底，科学数据资源超过了 150TB，提供在线服务的科学数据资源超过 100TB。数据资源涵盖物理、化学、地球科学、生物学、材料科学、能源科学、信息科学等多个学科领域；十二五期间的目标是形成开放共享、服务创新的国家级科技数据中心，为我国科技发展提供强大和持续的数据基础设施。此外，还开发了提供推送最新论文、专利和项目信息的科技信息服务，提供地理空间数据云等开放数据集，目前内容尚在完善过程中。

3.3 大数据发展现状分析

从国际上看，大数据方面的工作主要集中在以下 4 个方面：一是政府层面，主要是提供政策导向，推动政府数据、科学数据开放，为大数据发展提供政策支持和可信数据来源；二是研究机构利用政府资金，开展科学数据、论文等开放数据集建设，并开展数据集间互操作方面的研究；三是 Google 等公司研制了分布式数据处理平台等产品，为大数据发展提供技术和产品支撑；四是标准化方面，目前最为实质性的是 ISO/IECJTC1 成立了大数据研究组，由美国 NIST 牵头，NIST 系统地开展了大数据架构、数据、安全需求等方面的研究，研究成果将贡献至 JTC1。

从国内情况来看，多个地方政府提出大力发展大数据的政策导向，在北京市率先开放了政府数据资源；中国科学院计算机网络信息中心研发了科学数据库等开放数据集；阿里利用拥有的大量商业数据为基础，进行统计、分析和挖掘，对外提供数据服务；人民大学等研究院所和百度、阿里等公司正在开展大数据处理技术和平台研制工作；在标准化方面，全国信息技术标准化技术委员在充分调研

基础上，提出了技术体系参考模型和标准体系框架，提出了术语、体系结构、数据表示、非结构化数据、数据质量、科学数据集等方面标准，其中多项标准已经立项。

从大数据与相关技术的关联关系上来看，互联网、物联网、云计算等技术的发展为大数据提供了基础，互联网、物联网提供了大量数据来源；云计算的分布式存储和计算能力提供了技术支撑；而大数据的核心是数据处理。其中传统的数据处理技术经过演进依然有效，新兴技术还在不断探索和发展中。

从大数据商业模式上来看，大数据时代，不断涌现出围绕大数据、利用大数据的新产品形态、新业务模式。其中，“数据租售”即通过出售原始的业务数据或者是经过初步处理分析的数据来获取直接的利益，以商品化的数据应用创造了新的商业模式。百度游戏通过搜集整理网络游戏用户的搜索需求和搜索热点，建立完备的用户行为数据库，提供给上游的游戏运营商，创造了数据服务收入，成为在搜索引擎领域中将数据支持服务变为主要盈利模式的成功案例。阿里巴巴正在研发的数据仓库，以阿里巴巴拥有的大量商业数据为基础，进行统计、分析和挖掘，形成规范的实体明细数据和指标数据，对外服务。其中，“魔方”是淘宝网成立的专门用于提供数据服务的机构，为商家提供行业分析数据，从中获取利益。此外，科学机构、政府机构提供的数据集也成为可信的重要数据来源。

大数据的发展目前急需解决三方面的问题：一是提供处理大数据技术能力的平台；二是需要明确大数据生态环境中各个角色的权利、义务，解决数据开放和共享过程中的产权保护、权限管理和隐私保护等问题；三是需要建立可管理维护、可信、易于互操作的数据资源集，这是大数据发展的初步成果，为大数据处理、应用和进一步发展提供基础，也是我国的重要信息资源。其中第一个问题是技术问题，后面两个问题既是技术问题，也是管理问题。

4.大数据技术参考模型和关键技术

4.1 大数据技术参考模型

大数据作为一项新兴技术，目前尚未形成完善、达成共识的技术体系。本章结合 NIST 和 JTC1/SC32 的研究成果，结合我们对大数据的理解和分析，提出了大数据技术参考模型。

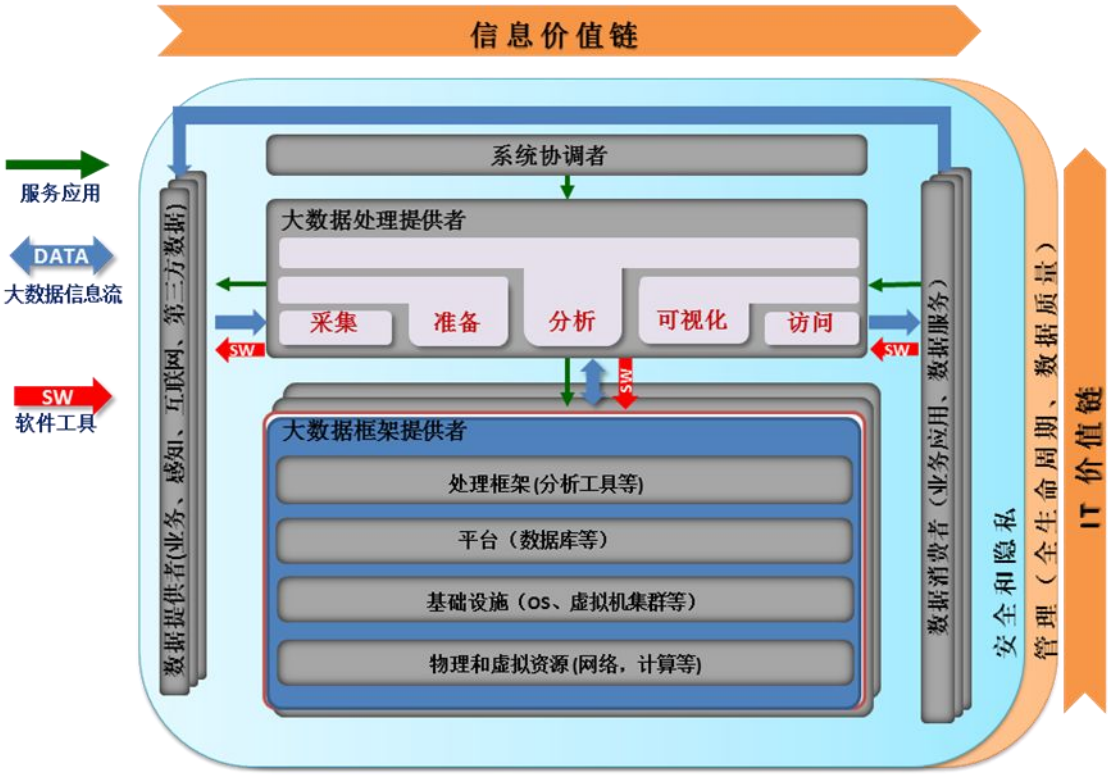

图 8 大数据技术参考模型图

大数据技术参考模型表示了通用的、技术无关的大数据系统的逻辑功能模块以及模块之间的互操作接口（如：服务）。这些被称为“提供者”的模块代表了大数据生态系统中的功能角色，表明他们提供或实施大数据系统中特定技术的功能。

大数据技术参考模型基于代表大数据价值链的两个维度组成：信息流（垂直维）和 IT 集成（水平维）。在信息流维度上，价值通过数据采集、集成、分析、使用结果来实现。在 IT 维度上，价值通过为大数据应用的实施提供拥有或运行

大数据的网络、基础设施、平台、应用工具以及其他 IT 服务来实现。大数据应用提供者模块是在两个维的交叉点上，表明大数据分析和其实施是为两个价值链上大数据利益相关者提供的特定价值。

五个主要的架构模块代表在每个大数据系统中存在的不同技术角色：数据提供者、数据消费者、大数据处理提供者、大数据框架提供者、系统协调者。另外两个架构模块是安全隐私和管理，代表能为大数据系统其他模块提供服务和功能的构件。这两个关键功能极其重要，因此也被集成在任何大数据解决方案中。

此架构可以用于多个大数据系统组成的复杂系统，这样其中一个系统的大数据使用者可以作为另外一个系统的大数据提供者。

“数据”箭头表明在系统主要模块之间流动的数据。模块之间的数据可以是物理实体或者数据的引用地址。“软件”箭头表明在大数据流程中的支撑软件工具。“服务使用”代表软件程序接口。虽然此架构主要用于实时运行环境，但也可用于配置阶段。大数据系统中涉及的人工协议和人工操作没有被包含在此架构中。

在数据提供者模块中，提供者应该包含业务、感知、互联网和第三方数据四个种类。其中业务数据提供者提供传统信息系统中存在并动态产生的大量的结构化数据和异构数据；感知数据提供者提供由物联感知设备实时生成的大量数据；互联网数据提供者提供由互联网应用快速生成的大量的非结构化数据；而第三方数据提供者则是提供政府、学术界、商业机构逐步对外开放了一些可维护管理、可信的数据集。

同数据提供者模块一样，对于数据消费者模块，我们将其分解为业务应用和数据服务平台。其中业务应用主要是根据各个行业不同的数据需求进行相应的处理，通过不同的方式，形成符合业务的应用。而数据服务平台则是通过平台向数据需求者提供相应的数据服务。

在管理模块中，我们认为大数据管理主要应该关注数据的全生命周期和数据质量的管理。数据质量贯穿数据全生命周期，涉及从采集、存储、传输、分析、展现等各方面。数据质量关键因素包括数据的一致性、数据的完整性、数据的准确性、数据的新鲜度、数据的约束性等。

另外我们认为数据提供者和数据消费者在不同角度下可以相互转换，数据消费者在处理使用数据之后，会形成新的数据提供出来。而数据提供者在收集整理相关数据的时候，也会变成数据消费者。因此，我们添加了一条由数据消费者指向数据提供者的数据流向，用以表达他们之间的可转化关系。

4.2 大数据核心技术

4.2.1 数据准备技术

1）数据表示

数据是用来描述对应用程序很重要的现实世界的信息资源。数据描述物、人、产品、项目、客户、资产和记录等，数据表示通过对于信息资源的分类、编码以及格式等内容进行分析规范，使得信息数据能够快速、高效、准确地被计算机所识别，从而使得采集上来的数据能够更好的为应用服务。表示数据的过程需要数据结构分析、文档准备和对等检查等。其最终结果是应用程序有关信息记录的概念性视图，回答数据“是什么、在哪里、何时以及为什么”等问题。数据表示的最终目标是将这些信息转化为计算机能够识别的语言，并存储起来。

2）元数据注册

元数据是描述对象的数据，用于说明对象的相关特征。对于某个对象的元数据描述，可以避免在不同环境、语境以及不同视角下对同一对象的差异化描述，确保对象描述的唯一性。大数据环境下由于数据获取与表现方式存在不同，所以数据类型有多种，包括声音、图像、视频、文本等。由于记录信息的角度不同、信息获取的方式不同，导致不同类型的数据在具体的数据处理和表现上也有着本质的差别。因此，在大量数据对于同一对象的不同描述中，如果能够提前对于该

对象的元数据进行注册，并在描述该对象时通过元数据的相关表示规范进行描述，可以有效的将对象内容中的唯一特性表示出来，从而为包含该对象的相关应用打下良好的基础。

3）本体元建模

源于哲学范畴的本体论（Ontology）在计算机科学技术领域，尤其是知识工程领域率先得到了应用。近年来,本体论已被广泛地应用于信息与知识的分类和表达领域,应用领域的本体得到了共享与重用。利用本体对应用领域相关知识进行建模能够有效地支撑信息的语义共享,本体及其形式化规范还能够应用于人-机通信、机-机通信与信息交换，有力地支撑系统的语义集成与互操作。本体描述语言 OWL(Web Ontology Language)和 OWL S(OWL for Web Service)等极大增强了本体的建模机制和表达能力，推动了语义服务计算的工程化进程。本体建模具有开放、伸缩地定义和描述语义关联的特性，从而具有随实际问题的语义丰简、智能化程度的需要，开放地表达与构造软件实体的语义行为能力。结合本体的建模理论与技术是大数据环境下，对于知识挖掘、信息潜在价值发现的重要技术支撑。

4.2.2 数据存储技术

1）分布式文件系统

分布式文件系统将大规模海量数据用文件的形式保存在不同的存储节点中，并用分布式系统进行管理。其技术特点是为了解决复杂问题，将大的任务分解为多个小任务，通过让多个处理器或多个计算机节点参与计算来解决问题。

分布式文件系统能够支持多台主机通过网络同时访问共享文件和存储目录，使多台计算机上的多个用户共享文件和存储资源。分布式文件系统架构更适用于互联网应用，能够更好地支持海量数据的存储和处理。基于新一代分布式计算的架构很可能成为未来主要的互联网计算架构之一。

目前典型的分布式文件系统产品有 GFS（Google File System 文件系统）、HDFS（Hadoop 分布式文件系统）等。

2）数据仓库

传统数据库并非专为数据分析而设计，数据仓库专用设备的兴起，表明面向事务性处理的传统数据库和面向分析的分析型数据库走向分离。

数据仓库专用设备，一般会采用软硬一体的方式。这类数据库采用更适于数据查询的技术，以列式存储或 MPP（大规模并行处理）技术为代表。数据仓库适合于存储关系复杂的数据模型（例如企业核心业务数据），适合进行一致性与事务性要求高的计算，以及复杂的 BI（商业智能）计算。在数据仓库中，经常使用数据温度技术、存储访问技术来提高性能。

a.列式存储

对于图像、视频、URL、地理位置等类型多样的数据，难以用传统的结构化方式描述，因此需要使用由多维表组成的面向列存储的数据管理系统来组织和管理数据。列式存储将数据按行排序，按列存储，将相同字段的数据作为一个列族来聚合存储。当只查询少数列族数据时，列式数据库可以减少读取数据量，减少数据装载和读入读出的时间，提高数据处理效率。按列存储还可以承载更大的数据量，获得高效的垂直数据压缩能力，降低数据存储开销。

b.数据温度技术

数据温度技术可以提高数据访问性能，区分经常被访问和很少被访问的数据。经常访问的是高温数据，这类数据存储在高速存储区，访问路径会非常直接，而低温数据则可以放在非高速存储区，访问路径也相对复杂。

c.存储访问技术

近两年，存储访问技术不断变化，例如固态硬盘数据仓库，用接近闪存的性能访问数据，比原来在磁盘上顺序读取数据快很多。内存数据库产品，在数据库

管理系统软件上进行优化，规避传统数据库（数据仓库）读取数据时的磁盘 I/O 操作，节省访问时间。

3）非关系型数据库技术（NoSql）相比传统关系型数据库，NoSQL 数据库发展的原因是数据作用域发生了改变，不再是整数和浮点等原始的数据类型，数据已经成为一个完整的文件。这对数据库技术提出了新的要求，要求能够对数据库进行高并发读写、高效率存储和访问，要求数据库具有高可扩展性和高可用性，并具有较低成本。NoSQL 使得数据库具备了非关系、可水平扩展、可分布和开源等特点，为非结构化数据管理提供支持。目前 NoSQL 数据库技术大多应用于互联网行业。

4.2.3 数据平台技术

1）面向服务的体系结构（SOA）

SOA（Service-oriented Architecture，面向服务的体系结构）是近年来软件规划和构建的一种新方法，以“服务”为基本元素和核心。最早由国际咨询机构 Gartner 公司于 1996 年提出，2003 年以后成为我国软件产业界关注的重点，并得到众多行业的广泛应用。SOA 是大数据的重要支撑技术，通过“服务”的方式支撑实现大数据的跨系统汇聚、共享、交换、分析、管理和访问。我国在 SOA 广泛应用实践的基础上推动了标准化工作，形成了支撑各类应用的服务技术架构系列标准，并在智慧城市、电子政务等众多信息化领域取得了成功实践，具备了支撑大数据发展的良好基础。

2）MapReduce 框架

MapReduce 是一个软件架构，用于大规模数据集（大于 1TB）的并行运算。MapReduce 框架是 Hadoop 的核心，但是除了 Hadoop，MapReduce 上还可以有 MPP（列数据库）或 NoSQL。

当处理一个大数据集查询时，MapReduce 会将任务分解并在运行的多个节点处理。当数据量很大时，一台服务器无法满足需求，分布式计算优势体现出来。

MapReduce 有将任务分发到多个服务器上处理大数据的能力。HDFS（Hadoop Distributed File System）的重要内容就是对于分布式计算，每个服务器都具备对数据的访问能力。

HDFS 与 MapReduce 的结合，使得在处理大数据的过程中计算性能得到保障。当 Hadoop 集群中的服务器出现错误时，整个计算过程不会终止；同时 HFDS 可保障在整个集群中发生故障错误时的数据冗余；当计算完成时将结果写入 HFDS 的一个节点之中。HDFS 对存储的数据格式并无苛刻的要求，数据可以是非结构化或其它类别。

Hadoop 是 MapReduce 框架的一个典型的应用。Hadoop 的可靠性是因为它假设计算元素和存储会失败，因此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；Hadoop 高效性是因为它以并行的方式工作，通过并行处理加快处理速度；Hadoop 还是可伸缩的，能够处理 PB 级数据。

4.2.4 数据处理技术

1）数据挖掘和分析

大数据只有通过分析才能获取很多智能的、深入的、有价值的信息。越来越多的应用涉及到大数据，而这些大数据的属性与特征，包括数量、速度、多样性等都是呈现了不断增长的复杂性，所以大数据的分析方法就显得尤为重要，可以说是数据资源是否具有价值的决定性因素。

大数据分析的理论核心就是数据挖掘，各种数据挖掘算法基于不同的数据类型和格式，可以更加科学地呈现出数据本身具备的特点，正是因为这些公认的统计方法使得深入数据内部、挖掘价值成为可能。另一方面，也是基于这些数据挖掘算法才能更快速的处理大数据。

大数据分析的使用者有大数据分析专家，同时还有普通用户，二者对于大数据分析最基本的要求是可视化。可视化分析能够直观地呈现大数据特点，同时能够非常容易被使用者所接受。

大数据分析离不开数据质量和数据管理，高质量的数据和有效的数据管理，无论是在学术研究还是在商业应用领域，都能够保证分析结果的真实和有价值。

数据挖掘和分析的相关方法如下：

——神经网络方法。神经网络具有良好的鲁棒性、自组织自适应性、并行处理、分布存储和高度容错等特性，非常适合解决数据挖掘的问题，用于分类、预测和模式识别的前馈式神经网络模型；以 hopfield 的离散模型和连续模型为代表，分别用于联想记忆和优化计算的反馈式神经网络模型；以 art 模型、koholon 模型为代表，用于聚类的自组织映射方法。神经网络方法的缺点是"黑箱"性，人们难以理解网络的学习和决策过程。

——遗传算法。遗传算法是一种基于生物自然选择与遗传机理的随机搜索算法，是一种仿生全局优化方法。遗传算法具有的隐含并行性、易于和其它模型结合等性质，使它在数据挖掘中被广泛应用。遗传算法的应用还体现在与神经网络、粗集等技术的结合上，如利用遗传算法优化神经网络结构，在不增加错误率的前提下，删除多余的连接和隐层单元；用遗传算法和 bp 算法结合训练神经网络，然后从网络提取规则等。

——决策树方法。决策树是一种常用于预测模型的算法，它通过将大量数据有目的地分类，从中找到一些有价值的、潜在的信息。主要优点是描述简单、分类速度快、特别适合大规模的数据处理。最有影响和最早的决策树方法是由 quinlan 提出的著名的基于信息熵的 id3 算法。

——粗集方法。粗集理论是一种研究不精确、不确定知识的数学工具。粗集方法有几个优点：不需要给出额外信息；简化输入信息的表达空间；算法简单，易于操作。粗集处理的对象是类似二维关系表的信息表。

——覆盖正例排斥反例方法。利用覆盖所有正例、排斥所有反例的思想来寻找规则。首先在正例集合中任选一个种子，到反例集合中逐个比较。与字段取值

构成的选择子相容则舍去，相反则保留。按此思想循环所有正例种子，将得到正例的规则(选择子的合取式)。

——统计分析方法。在数据库字段项之间存在两种关系：函数关系和相关关系，对它们的分析可采用统计学方法。对它们可进行常用统计分析、回归分析、相关分析、差异分析等。

——模糊集方法。利用模糊集合理论对实际问题进行模糊评判、模糊决策、模糊模式识别和模糊聚类分析。系统的复杂性越高，模糊性越强，一般模糊集合理论是用隶属度来刻画模糊事物的亦此亦彼性的。李德毅院士在传统模糊理论和概率统计的基础上，提出了定性定量不确定性转换模型--云模型，并形成了云理论。

2）内存计算

内存计算（In-Memory Computing），实质上是 CPU 直接从内存而非硬盘上读取数据，并对数据进行计算、分析。此项技术是对传统数据处理方式的一种加速，是实现商务智能中海量数据分析和实施数据分析的关键应用技术。

内存计算适合处理海量的数据，以及需要实时获得结果的数据。比如可以将一个企业近十年几乎所有的财务、营销、市场等各方面的数据一次性地保存在内存里，并在此基础上进行数据分析。当企业需要做快速的账务分析，或要对市场进行分析时，内存计算能够快速地按照需求完成。内存相对于磁盘，其读写速度要快很多倍。内存计算可以模拟一些数据分析的结果，实现对市场未来发展的预测，如需求性建模、航空天气预测、零售商品销量预测、产品定价策略等。

3)流处理技术

在大数据时代，数据的增长速度超过了存储容量的增长，在不远的将来，人们将无法存储所有的数据，同时，数据的价值会随着时间的流逝而不断减少，此外，很多数据涉及用户的隐私无法进行存储。对数据进行实时处理的流处理获得了人们越来越多的关注。

数据的实时处理是一个很有挑战性的工作，数据流本身具有持续达到、速度快且规模巨大等特点，因此通常不会对所有的数据进行永久化存储，而且数据环境处在不断的变化之中，系统很难准确掌握整个数据的全貌。由于响应时间的要求，流处理的过程基本在内存中完成，其处理方式更多地依赖于在内存中设计巧妙的概要数据结构(Synopsis data structure)，内存容量是限制流处理模型的一个主要瓶颈。以 PCM(相变存储器)为代表的 SCM(Storage Class Memory，储存级内存)设备的出现或许可以使内存未来不再成为流处理模型的制约。

数据流的理论及技术研究已经有十几年的历史，目前仍旧是研究热点。当前得到广泛应用的很多系统多数为支持分布式、并行处理的流处理系统，比较代表性的商用软件包括 IBM 的 StreamBase 和 InfoSphere Streams，开源系统则包括 Twitter 的 Storm、Yahoo 的 S4 等。

4.3 大数据安全与隐私

与当前其他的信息一样，大数据在存储、处理和传输等过程中面临安全风险，具有数据安全与隐私保护需求。而实现大数据安全与隐私保护，较以往其他安全问题更为棘手，因为，在大数据背景下，这些大数据运营商既是数据的生产者，又是数据的存储、管理者和使用者，因此，单纯通过技术手段限制商家对用户信息的使用，实现用户数据安全和隐私保护是极其困难的。大数据收集了各种来源、各种类型的数据，其中包含了很多和用户隐私相关的信息。大量事实表明，大数据未能妥善处理会对用户的隐私造成极大的侵害。很多时候人们有意识地将自己的行为隐藏起来，试图达到隐私保护的目的，但是，在大数据环境下，我们可以通过用户零散数据之间的关联属性，将某个人的很多行为数据聚集在一起时，他的隐私就很可能会暴露，因为有关他的信息已经足够多，这种隐性的数据暴露往往是个人无法预知和控制的。在大数据时代，人们面临的威胁并不仅限于个人隐私泄露，还在于基于大数据对人们状态和行为的预测。例如零售商可以通过历史纪录分析，得到顾客在衣食住行等方面的爱好、倾向等；社交网络分析研究也表明，可以通过其中的群组特性发现用户的属性，例如通过分析用户的微博等信息，可以发现用户的政治倾向、消费习惯以及其它爱好等。

我们需要对大数据中的用户数据和隐私进行保护。我们必须解决好大数据时代数据公开和数据安全和隐私保护之间的矛盾，如果仅仅因为担心数据安全和隐私问题而不公开数据，则大数据的价值无法体现，因此，大数据时代的隐私性主要体现在不暴露用户敏感信息的前提下进行有效的数据挖掘，这有别于传统的信息安全领域更加关注文件的私密性等安全属性。根据需要保护的内容不同，隐私保护又可以细分为位置隐私保护、标识符匿名保护和连接关系匿名保护等。但大数据时代的数据快速变化给隐私保护带来了新的挑战，因为现有隐私保护技术主要基于静态数据集，我们必须考虑如何在这种复杂环境下实现对动态数据的利用和隐私保护。当前很多组织都认识到了大数据的安全问题。并积极行动起来关注大数据安全问题。2012 年运安全联盟 CSA 组建了大数据工作组，旨在寻找针对大数据中的安全和隐私问题的解决方案。

5. 大数据标准体系

目前，大数据技术相关标准的研制还处于起步阶段，本部分对 ISO/IEC、ITU 等国际标准化组织、NIST、国内全国信标委已经开展的标准化工作进行梳理，依据大数据技术体系，从基础、技术、产品、应用等不同角度及进行分析，形成了大数据标准体系框架。对我国现有标准、在研标准和将提出的标准计划进行分析，形成了大数据标准体系。对于目前急需研制的标准进行了较为详细的分析，这部分将成为后续标准化工作的重点。

5.1ISO/IEC JTC1 SC32 大数据标准化工作情况

ISO/IEC JTC1 SC32 “数据管理和交换”分技术委员会,是与大数据关系最为密切的标准化组织。SC32 持续致力于研制信息系统环境内及之间的数据管理和交换标准，为跨行业领域协调数据管理能力提供技术性支持,其标准化技术内容涵盖:协调现有和新生数据标准化领域的参考模型和框架；负责数据域定义、数据类型和数据结构以及相关的语义等标准；负责用于持久存储、并发访问、并发更新和交换数据的语言、服务和协议等标准；负责用于构造、组织和注册元数据及共享和互操作相关的其他信息资源（电子商务等）的方法、语言服务和协议等标准。SC32 下设 4 个工作组和几个研究组，主要内容如下：

- WG1：电子业务

工作范围为：研制各组织使用的信息系统间全球互操作所需的开放电子数据交换方面的通用 IT 标准，包括商务和信息技术两方面的互操作标准。

- WG2：元数据

工作范围为：研制开发和维护有利于规范和管理的元数据、元模型和本体的标准，此类标准有助于理解和共享数据、信息和过程，支持互操作性，电子商务以及基于模型和基于服务的开发，包括：建议用于规定和管理元数据、元模型和本体的框架；规定和管理元数据、元模型和本体；规定和管理过程、服务和行为

数据；开发管理元数据、元模型和本体的机制，包括注册和存储；开发交换元数据、元模型和本体的机制，包括基于互联网、局域网等的语义。

- WG3：数据库语言

工作范围为：为动态规定、维护和描述多用户环境中的数据库结构和组件制定和维护语言标准；通过规定事务提交、恢复和安全机制提供额外的对数据库管理系统完整性的支持；为存储、访问和处理多并发用户使用的数据库结构中的数据制定和维护语言标准；为其他标准编程语言提供开发接口；为描述数据类型和行为的其他标准提供访问接口或为开发用户提供数据库组件。

- WG4：SQL 多媒体和应用包

工作范围为：规定各种应用领域使用的抽象数据类型包的定义。每一个抽象数据类型包的定义是使用数据库语言 SQL 标准中提供的用户定义类型机制来规定的，包括全文、空间、静态图像、静态图形、动画、视频、音频、地震和音乐等数据包。为支持用户根据应用 API 需求进行数据管理，其他包使用 SQL 语言机制来定义，而不是用户定义类型。

- 其他研究组

目前 SC32 还建立了下一代分析技术与大数据研究组（SG Next Generation Analytics and Big Data）、云计算元数据研究组（SG Metadata for Cloud Computing）和基于事实基础的建模元模型研究组（SG Metamodel for Fact Based Modelling）等专项研究组，这些研究组也在数据管理与交换领域进行了相关研究。

2012 年 SC32 在柏林全会上，决定成立下一代分析和大数据研究组。该研究组主要的研究内容为下一代分析、社会分析和底层技术支持领域中潜在的标准化需求。2013 年 SC32 庆州全会上，该研究组形成了正式的研究报告，并提交至 JTC1 审议。

该报告提供了关于大数据研究的抽象概念模型，并给出了大数据现有标准基础，包括元数据、数据存储和检索以及大数据所支持的复杂数据类型三个领域。报告还对大数据标准化的工作方向做出了说明，一是加强元数据标准研究；二是

加强数据存储标准研究；三是加强支持复杂的、半结构化和非结构化等数据类型标准研究；四是深入研究 SC32 相关标准，做好协调工作。通过以上四个方向的探索实践，形成有效的标准化成果，用以支持下一代分析和大数据的相关项目开展和工具应用。

2014 年 6 月 SC32 在北京全会上，批准了 4 项为大数据提供标准化支持的新工作项：SQL 对多维数组的支持、SQL 对 JSON 的支持、数据集注册元模型和数据源注册元模型。其中 SQL 对 JSON 的支持由中国专家担任编辑。

SC32 在 2014 年北京全会期间举办了主题为“大数据标准化”的开放论坛，为国内外大数据领域的专家学者和产业管理部门人员、IT 界的骨干企业提供了一个开放交流的平台。来自于国内外大数据研究、应用及服务提供领域的专家学者做了相关主题报告，展现了当前大数据技术与标准的发展和应用前景。

JTC1/SC32 现有的标准制定和研究工作为大数据的发展提供了良好基础。

5.2 ISO/IEC JTC1 SG2 大数据标准化工作情况

ISO/IEC JTC1 于 2013 年 11 月全会上新成立负责大数据国际标准化的大数据研究组（ISO/IEC JTC1 SG2）。美国国家标准与技术研究院（NIST）专家 Wo Chang 担任召集人。

2014 年 ISO/IEC JTC1 SG2 的工作重点包括：调研国际标准化组织（ISO）、国际电工委员会（IEC）、第 1 联合技术委员会（ISO/IEC JTC1）等在大数据领域的关键技术、参考模型以及用例等标准基础；确定大数据领域应用需要的术语与定义；评估分析当前大数据标准的具体需求，提出 ISO/IEC JTC1 大数据标准优先顺序；向 2014 年 ISO/IEC JTC1 全会提交大数据建议的技术报告和其他研究成果。

大数据研究组的成立，标志着 JTC 1 统筹开展大数据的标准化工作，有利于大数据国际、国内标准化工作的开展。ISO/IEC JTC1 SG2 于 2014 年计划召开四次会议，前三次会议的形式为会议全程四天，前两天为成果交流展示，后两天为

具体标准工作讨论。第一次会议于 2014 年 3 月 18 日至 21 日在美国圣地亚哥超级计算中心召开；第二次会议于 2014 年 5 月 13 日至 16 日在荷兰阿姆斯特丹大学召开。第三次会议于 2014 年 6 月 16 日至 19 日在中国北京航空航天大学召开。第四次会议计划于 2014 年 9 月在英国召开。

5.3ITU 大数据标准化工作情况

ITU 在 2013 年 11 月发布了题目为《大数据:今天巨大，明天平常》的技术观察报告，这个技术观察报告分析了大数据相关的应用实例，指出大数据的基本特征、促进大数据发展的技术，在报告的最后部分分析了大数据面临的挑战和 ITU-T 可能开展的标准化工作。在这份报告中，特别提及了 NIST 和 JTC1/SC32 正在开展的工作。

从 ITU-T 的角度来看，大数据发展面临的最大挑战包括：数据保护、隐私和网络安全；法律和法规的完善。根据 ITU-T 现有的工作基础，开展的标准化工作包括：高吞吐量、低延迟、安全、灵活和规模化的网络基础设施；汇聚数据机和匿名；网络数据分析；垂直行业平台的互操作；多媒体分析；开放数据标准。ITU-T 正在开展的工作中，与大数据最为密切相关的是提出了一项题为“基于大数据的云计算的需求和能力”的新工作项目，以来自中国、韩国和波兰的专家为主研制。

5.4 NIST 标准化工作情况

NIST 建立的大数据公共工作组（NBD-PWG），工作范围是建立来自于产业界、学术界和政府的公共环境，共同形成达成共识的定义、术语、安全参考体系结构和技术路线图，提出数据分析技术应满足的互操作、可移植性、可用性和扩展性需求和安全有效地支持大数据应用的技术基础设施，用于为大数据相关方选择最佳的方案。

NBD-PWG 是一个开放工作组，欢迎来自于产业界、学术界和政府的各方面力量参与并贡献力量。原则上，工作组每周召开一次会议。工作组下设术语和定义、用例和需求、安全和隐私、参考体系结构和技术路线图 5 个分组，目前正在研制《大数据定义》、《大数据术语》、《大数据需求》、《大数据安全和隐私需求》、

《大数据参考体系结构》和《大数据技术路线图》等输出物，均已经形成了初步版本。

5.5 国内标准化工作情况

全国信息技术标准化技术委员会（TC28）[5]持续开展数据标准化工作,在元数据、数据库、数据建模、数据交换与管理等领域推动相关标准的研制与应用，为提升跨行业领域数据管理能力提供标准化支持。全国信标委中与大数据关系比较密切的组织包括：信标委非结构化数据管理标准工作组、信标委云计算工作组、信标委 SOA 分技术委员会、信标委传感器网络工作组等。此外大数据安全部分的标准与全国信息安全标准化技术委员会密切相关。

全国信标委于 2012 年成立了非结构化数据管理标准工作组，对口 ISO/IEC JTC1 SC32 WG4。非机构化数据管理标准工作组联合产、学、研、用等力量，致力于制定非结构化数据管理体系结构、数据模型、查询语言、数据挖掘、信息集成、信息提取、应用模式等相关国家标准和行业标准。目前正在开展《非结构化数据表示规范》、《非结构化数据访问接口规范》、《非结构化数据管理系统技术要求》等国家标准研制。

全国信标委云计算标准工作组目前正在开展大数据存储和分析应用的研究工作，旨在研究大数据存储和分析技术的应用分析、技术框架和标准研究等。同时，正在组织编制《云数据存储和管理》系列国家标准，为推动大数据存储和分析标准研究奠定了基础。

全国信标委 SOA 分技术委员会（以下简称“SOA 分委会”）负责面向服务的体系结构（SOA）、Web 服务和中间件的专业标准化的技术归口工作，并协助全国信息技术标准化技术委员会承担国际标准化组织相应分技术委员会的国内归口工作，现有成员 108 家。SOA 分委会还同时负责推动软件构件、云计算技术、智慧城市领域的标准化工作。2013 年 7 月 5 日，SOA 分委会全会上决定在基础

[5]全国信息技术标准化技术委员会简称全国信标委（TC 28）

工作组内启动大数据预研项目，目前正在征集成员阶段；2013 年 7 月 22 日开展了《大数据应用、技术、产业与标准化调研》，作为下一步大数据标准化研究的基础；此外，SOA 分委会智慧城市应用工作组在推动智慧城市中大数据的应用和服务化的标准研究。

全国信息安全标准化委员会（TC260）[6]是在信息安全技术专业领域内，从事信息安全标准化工作的技术工作组织。委员会负责组织开展国内信息安全有关的标准化技术工作，技术委员会主要工作范围包括：安全技术、安全机制、安全服务、安全管理、安全评估等领域的标准化技术工作。全国信安标委目前正开展大数据安全技术、产业和标准研究，为大数据的安全保障提供支撑。

5.6 大数据标准体系框架

在研究提出大数据技术框架的基础上，结合数据全周期管理，数据自身标准化特点，当前各领域推动大数据应用的初步实践，以及未来大数据发展的趋势，我们提出了大数据标准体系框架，如下图所示。

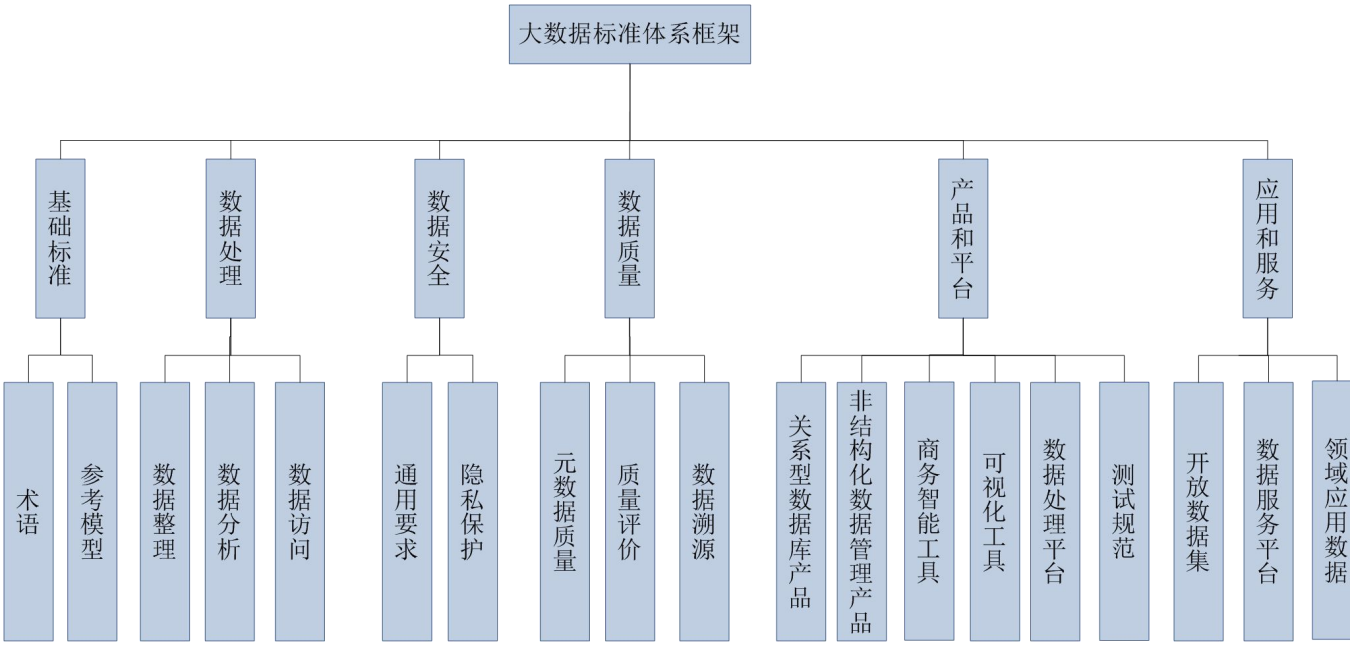

图 9 大数据标准体系框架

[6]全国信息安全标准化技术委员会简称全国信安标委（TC 260）

大数据标准体系由六个类别的标准组成，分别为：基础标准，数据处理标准，数据安全标准，数据质量标准，产品和平台标准及应用和服务标准。

a）基础标准

为整个标准体系提供包括总则、术语和参考模型等基础性标准。

b）数据处理标准

数据处理类标准包含数据整理、数据分析和数据访问三种类型的标准。

数据整理标准主要是针对数据在采集汇聚后的初步处理方式、方法的标准，包括数据表示、数据注册和数据清理三类标准。数据分析标准主要针对大数据环境下数据分析的性能、功能等要求进行规范。数据访问标准则是提供标准化的接口和共享方式，使数据能够被广泛的应用。

c）数据安全标准

数据安全作为数据标准的支撑体系，贯穿于数据整个生命周期的各个阶段。抛开传统的网络安全和系统安全，大数据时代下的数据安全标准主要包括通用要求、隐私保护两类标准。

d）数据质量标准

该类标准主要针对数据质量提出具体的管理要求和相应的指标要求，确保数据在产生、存储、交换和使用等各个环节中的质量，为大数据应用打下良好的基础。并对数据全生命周期进行规范化管理。主要包括元数据质量、质量评价和数据溯源三类标准。

e）产品和平台标准

该类标准主要针对大数据相关技术产品和应用平台进行规范。包括关系型数据库产品、非结构化数据管理产品、商务智能工具、可视化工具、数据处理平台和测试规范六类标准。其中关系型数据库产品标准针对存储和处理大数据的关系

型数据库管理系统，从访问接口、技术要求、测试要求等方面进行规范，为关系型数据库管理系统进行大数据的高端事务处理和海量数据分析提供支持；非结构化数据管理产品标准针对存储和处理大数据的非结构化数据管理系统，从参考架构、数据表示、访问接口、技术要求、测试要求等方面进行规范；商务智能工具用来帮助用户对大数据进行分析决策，包括 ETL、OLAP、数据挖掘等工具，商务智能工具标准对商务智能工具的技术及功能进行规范；可视化工具是对大数据处理应用过程中所需用到的可视化展现工具的技术和功能要求进行规范；数据处理平台标准是针对大数据处理平台从技术架构、建设方案、平台接口等方面进行规范；测试规范针对处理大数据的产品和平台给出测试方法和要求。

f）应用和服务标准

应用和服务类标准主要是针对大数据所能提供的应用和服务从技术、功能、开发、维护和管理等方面进行规范。主要包括开放数据集、数据服务平台和领域应用数据三类标准。其中开放数据集标准主要对向第三方提供的开放数据包中的内容、格式等进行规范；数据服务平台标准是针对大数据服务平台所提出的功能性、维护性和管理性的标准；领域应用数据指的是各领域根据其领域特性产生的专用数据标准。

5.7 大数据标准体系表

大数据标准体系表如下：

序号	一级分类	二级分类	国家标准编号	标准名称	采用标准号及采用程度	状态
1.	基础标准	总则		信息技术大数据标准化指南		拟研制
2.		术语		信息技术大数据术语		已申报
3.		参考模型		信息技术大数据参考模型		已申报
4.	数据处理	数据整理	GB/T18142-2000	信息技术数据元素值格式记法	idt ISO/IEC14957：1996	已发布
5.			GB/T18391.1-2009	信息技术元数据注册系统(MDR) 第 1 部分：框架	ISO/IEC11179-1：2004，	已发布

					IDT	
6.			GB/T18391.2-2009	信息技术元数据注册系统(MDR) 第 2 部分：分类	ISO/IEC11179-2：2005，IDT	已发布
7.			GB/T18391.3-2009	信息技术元数据注册系统(MDR) 第 3 部分：注册系统元模型与基本属性	ISO/IEC11179-3：2003，IDT	已发布
8.			GB/T18391.4-2009	信息技术元数据注册系统(MDR) 第 4 部分：数据定义的形成	ISO/IEC11179-4：2004，IDT	已发布
9.			GB/T18391.5-2009	信息技术元数据注册系统(MDR) 第 5 部分：命名和标识原则	ISO/IEC11179-5：2005，IDT	已发布
10.			GB/T18391.6-2009	信息技术元数据注册系统(MDR) 第 6 部分：注册	ISO/IEC11179-6：2005，IDT	已发布
11.			GB/T 21025-2007	XML 使用指南		已发布
12.			GB/T 23824.1-2009	信息技术实现元数据注册系统内容一致性的规程第 1 部分：数据元	ISO/IEC TR20943-1：2003，IDT	已发布
13.			GB/T 23824.3-2009	信息技术实现元数据注册系统内容一致性的规程第 3 部分：值域	ISO/IEC TR20943-3：2004，IDT	已发布
14.			20051294-T-339	信息技术元模型互操作性框架第 1 部分：参考模型		已报批
15.			20051295-T-339	信息技术元模型互操作性框架第 2 部分：核心模型		已报批
16.			20051296-T-339	信息技术元模型互操作性框架第 3 部分：本体注册的元模型		已报批
17.			20051297-T-339	信息技术元模型互操作性框架第 4 部分：模型映射的元模型		已报批
18.			20080046-T-469	信息技术元数据模块(MM) 第 1 部分:框架	ISO/IEC19773-1:2011	已报批
19.			20080044-T-469	信息技术技术标准及规范文件的元数据		已报批
20.			20080045-T-469	信息技术通用逻辑基于逻辑的语系的框架	ISO/IEC 24706:2005	已报批
21.			20080485-T-469	跨平台的元数据检索、提取与汇交协议		已报批
22.				信息技术异构媒体数据统一语义描述		已申报

23.		**数据分析**		信息技术大数据分析总体技术要求		拟研制
24.				信息技术大数据分析过程模型参考指南		拟研制
25.		**数据访问**	GB/T 12991-2008	信息技术数据库语言 SQL 第 1 部分：框架	ISO/IEC9075-1：2003，IDT	已发布
26.			20120567-T-469	信息技术云数据存储和管理第 1 部分：总则		在研
27.			20120568-T-469	信息技术云数据存储和管理第 2 部分：基于对象的云存储应用接口		在研
28.			20120569-T-469	信息技术云数据存储和管理第 5 部分：基于 Key-Value 的云数据管理应用接口		在研
29.				信息技术通用数据导入接口规范		已申报
30.				信息技术通用数据导入接口测试规范		拟研制
31.	**数据安全**	**通用要求**	GB/T 20009-2005	信息安全技术数据库管理系统安全评估准则		已发布
32.			GB/T 20273-2006	信息安全技术数据库管理系统安全技术要求		已发布
33.			GB/T 22080-2008	信息技术安全技术信息安全管理体系要求		已发布
34.			GB/T 22081-2008	信息技术安全技术信息安全管理实用规则		已发布
35.			20100383-T-469	信息技术安全技术信息安全管理体系实施指南		已发布
36.				信息安全技术数据库管理系统安全技术要求		已立项
37.				信息安全技术信息技术产品在线服务信息安全规范		已立项
38.				信息安全技术云计算服务安全能力要求		已立项
39.				信息安全技术大数据安全指南		拟研制
40.				信息安全技术大数据安全参考架构		拟研制
41.				信息安全技术大数据全生命周期安全要求		拟研制
42.		**隐私保护**	GB/Z 28828-2012	信息安全技术公共及商用服务信息系统个人信息保护指南		已发布
43.			20130323-T-469	信息安全技术个人信息保护管理要求		在研
44.			20130338-T-469	信息安全技术移动智能终端个人信息保护技术要求		在研
45.				信息安全技术个人信息保护指南		已立项
46.				信息安全技术大数据中的隐私保护规范		拟研制
47.	**数据质量**	**元数据质**	2010-3324T-SJ	信息技术元数据质量要求框架		在研
48.			2010-3325T-SJ	信息技术元数据质量指标		在研

		量				
49.		**质量评价**		软件工程软件产品质量要求和评价（SQuaRE）数据质量模型		已立项
50.				数据能力成熟度模型规范		已申报
51.				信息技术数据质量评价指标		拟研制
52.		**数据溯源**		信息技术数据引用规范		拟研制
53.				信息技术数据溯源描述模型		拟研制
54.	**产品和平台**	**关系型数据库产品**	GB/T 28821-1012	关系数据管理系统技术要求		已发布
55.			20080484-T-469	关系数据库管理系统检测规范		已报批
56.			20100401-T-469	分布式关系数据库服务接口规范		在研
57.		**非结构化数据管理产品**	20121409-T-469	非结构化数据表示规范		在研
58.			20121410-T-469	非结构化数据访问接口规范		在研
59.			20121411-T-469	非结构化数据管理系统技术要求		在研
60.				实时数据库通用接口规范		已申报
61.				非结构化数据管理系统参考模型		已申报
62.				非结构化数据管理术语		拟研制
63.				非结构化数据查询语言		拟研制
64.		**可视化工具**		大数据可视化工具通用要求		拟研制
65.		**数据处理平台**		大数据平台通用数据存储结构规范		拟研制
66.				大数据平台通用软件开发工具包（SDK）规范		拟研制
67.	**应用和服务**	**开放数据集**		开放数据集基本要求		拟研制
68.				开放数据集标识管理		拟研制
69.		**数据服务平台**	GB/T 29262-2012	信息技术面向服务的体系结构（SOA）术语		已发布
70.			GB/T 29263-2012	信息技术面向服务的体系结构（SOA）应用的总体技术要求		已发布
71.				信息技术 数据交易服务平台 通用功能要求		已申报
72.				信息技术数据交易平台 交易数据描述		已申报
73.				数据服务平台管理操作规程		拟研制

74.		领域应用数据				
75.						

根据大数据标准体系框架，整理出已发布、已报批、已立项、已申报、在研以及拟研制的大数据相关国家标准 73 项。

通过对现有各类标准情况进行分析可以看出：

（1） 从技术标准上来看，大数据相关的技术标准具有一定的工作基础。在数据整理方面，我国已经研制的一些相关标准，同样适用于大数据环境，目前急需加强这类标准的推广应用；数据分析是大数据的特点和难点，标准较为缺乏；在数据访问方面，目前在研多项数据库、云数据存储和管理类标准，适用于大数据底层数据接口，但是尚缺乏数据导入、导出类标准；数据安全方面，部分现有标准适用，但是尚缺乏针对大数据的安全框架、隐私、访问控制类标准；数据质量是大数据应用和发展的基础，目前有多项在研标准，但是均尚未发布，较为缺乏；

（2） 针对大数据产品和平台，目前在研多项数据库、非结构化数据管理产品类标准，尚无针对大数据可视化工具、数据处理平台的标准；在大数据环境下，数据也已成为产品，而针对开放数据集、数据服务平台等新兴产品和服务形态，尚缺乏相应的标准。

综上所述，针对大数据，我国在数据管理、云计算、信息安全等方面，已经发布和在研一些标准，适用于大数据环境，提供了一定的基础，但是缺乏标准化整体规划；数据分析、数据安全、数据质量管理等技术标准，数据处理平台、开放数据集、数据服务平台类新型产品和服务形态的标准较为缺乏，急需研制。

5.8 近期急需研制标准

结合当前对大数据技术与产业发展需求的分析，根据开展的大数据技术、应用、产业与标准化调研工作，提出近期急需重点研制的标准项目。如下所示：

序号	一级分类	二级分类	国家标准编号	标准名称	状态
1	基础标准	总则		信息技术大数据标准化指南	拟研制
2		术语		信息技术大数据术语	已申报
3		参考模型		信息技术大数据参考模型	已申报
4	数据处理	数据整理		信息技术异构媒体数据统一语义描述	已申报
5		数据分析		信息技术大数据分析总体技术要求	拟研制
6				信息技术大数据分析过程模型参考指南	拟研制
7		数据访问	20120567-T-469	信息技术云数据存储和管理第 1 部分：总则	在研
8			20120568-T-469	信息技术云数据存储和管理第 2 部分：基于对象的云存储应用接口	在研
9			20120569-T-469	信息技术云数据存储和管理第 5 部分：基于 Key-Value 的云数据管理应用接口	在研
10				信息技术通用数据导入接口规范	已申报
11				信息技术通用数据导入接口测试规范	拟研制
12	数据安全	通用要求		信息安全技术数据库管理系统安全技术要求	拟研制
13				信息安全技术信息技术产品在线服务信息安全规范	拟研制
14				信息安全技术云计算服务安全能力要求	拟研制
15				信息安全技术大数据安全指南	拟研制
16				信息安全技术大数据安全参考架构	拟研制
17				信息安全技术大数据全生命周期安全要求	拟研制
18		隐私保护	20130323-T-469	信息安全技术个人信息保护管理要求	在研
19			20130338-T-469	信息安全技术移动智能终端个人信息保护技术要求	在研
20				信息安全技术个人信息保护指南	拟研制
21				信息安全技术大数据中的隐私保护规范	拟研制
22	数据质量	元数据质量	2010-3324T-SJ	信息技术元数据质量要求框架	在研
23			2010-3325T-SJ	信息技术元数据质量指标	在研
24		质量评价		软件工程软件产品质量要求和评价（SQuaRE）数据质量模型	已立项
25				数据能力成熟度模型规范	已申报
26				信息技术数据质量评价指标	拟研制
27		数据溯源		信息技术数据引用规范	拟研制
28				信息技术数据溯源描述模型	拟研制
29	产品和平台	关系型数据库产品	20080484-T-469	关系数据库管理系统检测规范	已报批
30			20100401-T-469	分布式关系数据库服务接口规范	在研
31		非结构化数据管理产品	20121409-T-469	非结构化数据表示规范	在研
32			20121410-T-469	非结构化数据访问接口规范	在研
33			20121411-T-469	非结构化数据管理系统技术要求	在研
34				实时数据库通用接口规范	已申报
35				非结构化数据管理系统参考模型	已申报
36				非结构化数据管理术语	拟研制

37				非结构化数据查询语言	拟研制
38		**可视化工具**		大数据可视化工具通用要求	拟研制
41		**数据处理平台**		大数据平台通用数据存储结构规范	拟研制
42				大数据平台通用软件开发工具包（SDK）规范	拟研制
39	**数据服务平台**	**开放数据集**		开放数据集基本要求	拟研制
40				开放数据集标识要求	拟研制
43		**数据服务平台**		信息技术数据交易平台 交易数据描述	已申报
44				信息技术 数据交易服务平台 通用功能要求	已申报
45				数据服务平台管理操作规程	拟研制

6. 我国大数据工作重点建议

通过对调研数据的分析，我们发现“政府的统一规划和指导”、“相关标准的制定”以及“相关技术的成熟”被大多数受访者认为是推动大数据相关产业发展的关键因素（见图 10），结合调研结果，本报告给出了我国大数据工作重点建议。

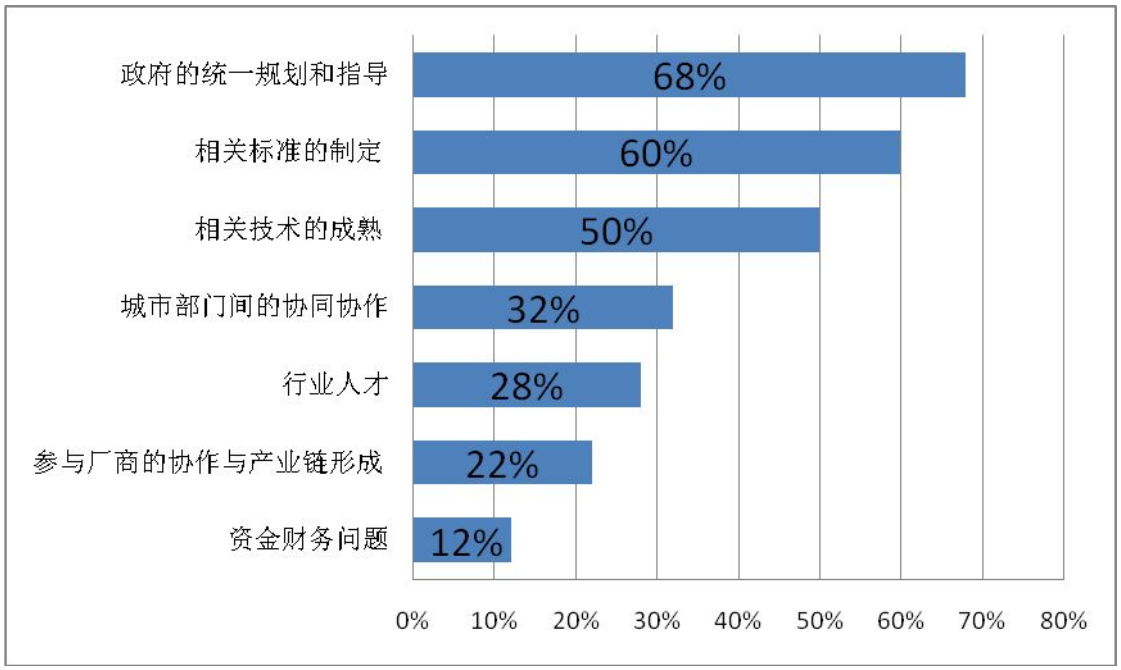

图 10 推动大数据相关产业发展的关键因素

6.1 加强大数据相关政策法规研究

大数据是一项新兴技术，将带来产业结构和管理方式的改变，同时各类数据本身也是国家重要的基础信息资源。因此建议从国家层面，系统地开展大数据相关方权利和义务、大数据各个业务环节基本操作规程、数据内容保护、个人隐私保护等各方面政策法规研究，保障大数据内容安全可控，产业和应用规范发展。

6.2 加强大数据核心技术研究

目前，大数据应用需求强烈，需要加强数据表示、数据分析、数据存储、数据质量、数据安全体系、大数据集互操作性等大数据核心技术的研究，才能为大数据应用提供技术保障。

6.3 推动开放数据集建设

开放数据集是大数据发展重要的数据来源之一，也是我国重要的信息资源。因此建议从国家层面，系统地开展安全可控的开放数据集建设。从科学数据、科学论文等公共资源入手，建设科学领域开放大数据集；从国家层面，逐步建立分层管理、自主可控的工业产品数据集；鼓励商业企业开放原始数据或处理后数据；在科学数据、工业数据、商业数据研究的基础上，逐步探索政府开放平台等经验，加强运行机制研究与建设，加快推进政府信息资源共享；循序渐进建立特定主题的数据监测系统，如在交通、能源、医疗、自然灾害等专题建立基础数据库，加强大数据集间互操作性研究。为大数据发展提供基础，推动国家基础数据安全可控、开放共享，促进大数据成果广泛应用。

6.4 鼓励利用大数据开展服务，创新大数据应用模式

建议国家研究出台相关政策，鼓励各类机构利用大数据提供公共服务，形成数据再开发利用的良性循环机制。调动企业的积极性，鼓励企业创新大数据的应用模式，利用大数据提高企业本身分析和决策能力，提高对外服务能力。

6.5 系统开展大数据标准化工作

大数据标准化工作是支撑大数据产业发展和应用的重要基础。

目前国际、国内大数据标准化工作都刚刚起步，建议尽快成立大数据标准化工作组，吸纳国内产学研用各方面的力量，国内、国际标准化工作同步发展，梳理国内各方面成果，系统地研究国际先进成果，适时参与国际标准化工作。

建议加强大数据标准化顶层设计，从产品、技术、安全、管理、应用等多个角度梳理大数据标准需求，认真分析智慧城市、云计算、移动互联网等相关领域与大数据的关系，建立健全大数据标准体系，重点突破一批涉及大数据发展的基础性、方法性、公共性标准的研制，为大数据发展和应用夯实标准化基础。

7.参考文献

[1]何克清，何非，李兵，等. 面向服务的本体元建模理论与方法研究[J]. 计算机学报, 2005, 4.

[2] Howe D, Costanzo M, Fey P, et al. Big data: The future of biocuration[J]. Nature, 2008, 455(7209): 47-50.

[3] Lynch C. Big data: How do your data grow?[J]. Nature, 2008, 455(7209): 28-29.

[4] Lohr S. The age of big data[J]. New York Times, 2012, 11.

[5]Bughin J, Chui M, Manyika J. Clouds, big data, and smart assets: Ten tech-enabled business trends to watch[J]. McKinsey Quarterly, 2010, 56.

[6] 王珊，王会举，覃雄派，等. 架构大数据：挑战，现状与展望[J]. 计算机学报, 2011, 34(10): 1741-1752.

[7] 李国杰. 大数据研究的科学价值 [J][J]. 中国计算机学会通讯, 2012, 8(9): 8-15.

[8] 姜奇平. 大数据时代到来[J]. 互联网周刊, 2012 (2): 6-6.

[9]Mayer-Schonberger V, Cukier K N. Big Data: A Revolution That Will Transform How We Live, Work, and Think[M]. Eamon Dolan/Houghton Mifflin Harcourt, 2013.

[10] Zikopoulos P, Eaton C. Understanding big data: Analytics for enterprise class hadoop and streaming data[M]. McGraw-Hill Osborne Media, 2011.

附件调研单位名单（共 74 家）

高校及科研单位	
北京航空航天大学	人民大学信息学院
北京大学	华南理工大学软件学院
中科院软件所	华南师范大学计算机学院
工业和信息化部电子第五研究所	上海软件产业促进中心
复旦大学	华北计算技术研究所
东北大学	华中科技大学电子与信息工程系
北京邮电大学	中国科学院高能物理研究所计算中心
中科院上海微系统与信息技术研究所	清华大学软件学院
北京交通大学	广西师范大学计算机科学与信息工程学院
广西大学计算机与电子信息学院	浙江省标准化研究院
上海计算机软件技术开发中心	西安电子科技大学
武汉大学软件工程国家重点实验室	上海浦东智慧城市发展研究院
国防科大计算机学院	北京科技大学计算机与通信工程学院
西安交通大学	中国电子科技集团公司第二十八研究所
企业	
北京东方通科技发展有限责任公司	Oracle（中国）有限公司
中创软件商用中间件股份有限公司	国际商业机器中国（投资）有限公司
上海普元信息技术股份有限公司	SAP（中国）有限公司
北京中软国际信息技术有限公司	微软（中国）有限公司
山东浪潮齐鲁软件产业股份有限公司	北京汇金科技股份有限公司
长风开放标准平台软件联盟	阿里云计算有限公司
深圳市金蝶中间件有限公司	普华基础软件股份有限公司
上海宝信软件股份有限公司	神州数码信息系统有限公司
太极计算机股份有限公司	腾讯科技（深圳）有限公司
华迪计算机集团有限公司	北京三星通信技术研究有限公司
北京锐易特软件技术有限公司	华为技术有限公司
北京炎黄盈动科技有限公司	方正宽带网络股份有限公司
用友软件股份有限公司	上海互联网软件有限公司
北京有生博大软件技术有限公司	北京中油瑞飞信息技术有限责任公司
北京希艾欧管理技术有限公司	北京赛迪时代信息产业股份有限公司
南京朗坤软件有限公司	科大国创软件股份有限公司
大唐软件技术股份有限公司	中兴通讯股份有限公司
北京开普互联科技有限公司	北京华胜天成科技股份有限公司
东软集团股份有限公司	北京市闪联信息产业协会
启明信息技术股份有限公司	北京控三兴信息技术有限公司
中国软件与技术服务股份有限公司	南京斯坦德通信股份有限公司
石化盈科信息技术有限责任公司	希捷国际科技（无锡）有限公司
万国数据服务有限公司	华数数字电视传媒集团有限公司