CAICT 中国信通院

大 数 据 白 皮 书

（2019年）

中国信息通信研究院

2019年12月

版权声明

本白皮书版权属于中国信息通信研究院，并受法律保护。转载、摘编或利用其它方式使用本白皮书文字或者观点的，应注明“来源：中国信息通信研究院”。违反上述声明者，本院将追究其相关法律责任。

前　言

当前，全球大数据正进入加速发展时期，技术产业与应用创新不断迈向新高度。大数据通过数字化丰富要素供给，通过网络化扩大组织边界，通过智能化提升产出效能，不仅是推进网络强国建设的重要领域，更是新时代加快实体经济质量变革、效率变革、动力变革的战略依托。

本白皮书是继《大数据白皮书（2014 年）》、《大数据白皮书（2016 年）》、《大数据白皮书（2018 年）》之后中国信通院第四次发布大数据白皮书。本白皮书在前三版的基础上，聚焦一年多来大数据各领域的进展和趋势，梳理主要问题并进行展望。在技术方面，重点探讨了近两年最新的大数据技术及其融合发展趋势；在产业方面，重点讨论了我国大数据产品的发展情况；在数据资产管理方面，介绍了行业数据资产管理、数据资产管理工具的最新发展情况，并着重探讨了数据资产化的关键问题；在安全方面，从多种角度分析了大数据面临的安全问题和技术工具。希望本白皮书的分析可以对政府和行业提供参考。

目 录

CAICT 中国信通院

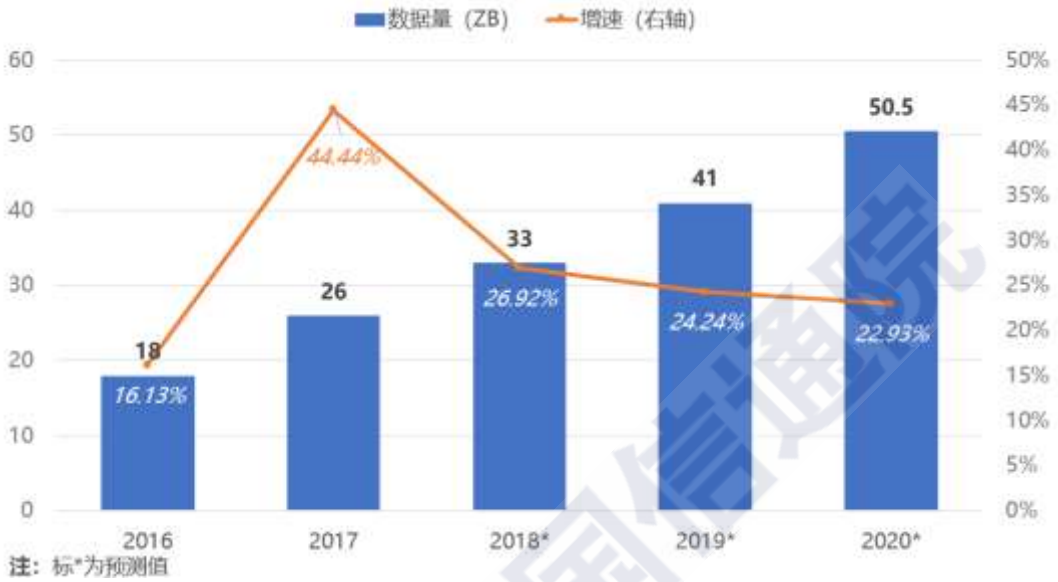
图 表 目 录

CAICT 中国信通院

CAICT 中国信通院

一、国际大数据发展概述

近年来，全球大数据的发展仍处于活跃阶段。根据国际权威机构 Statista 的统计和预测，全球数据量在 2019 年有望达到 41ZB[1]。



数据来源：IDC、Seagate、Statista estimates

图 1　全球每年产生数据量估算图

2019 年以来，全球大数据技术、产业、应用等多方面的发展呈现了新的趋势，也正在进入新的阶段。本章将对国外大数据战略、技术、产业等领域的最新进展进行简要叙述。

（一）大数据战略持续拓展

相对于几年前，2019 年国外大数据发展在政策方面略显平淡，只有美国的《联邦数据战略第一年度行动计划（Federal Data Strategy Year-1 Action Plan）》草案比较受到关注。

2019 年 6 月 5 日，美国发布了《联邦数据战略第一年度行动计划》草案，这个草案包含了每个机构开展工作的具体可交付成果，以

[1] ZB，即十万亿亿字节，相当于 2^{40}GB

及由多个机构共同协作推动的政府行动，旨在编纂联邦机构如何利用计划、统计和任务支持数据作为战略资产来发展经济、提高联邦政府的效率、促进监督和提高透明度[2]。

相对于三年前颁布的《联邦大数据研发战略计划》，美国对于数据的重视程度继续提升，并出现了聚焦点从“技术”到“资产”的转变，其中更是着重提到了金融数据和地理信息数据的标准统一问题。此外，配套文件中“共享行动：政府范围内的数据服务”成为亮点，针对数据跨机构协同与共享，从执行机构到时间节点都进行了战略部署。

早些时候，欧洲议会通过了一项决议，敦促欧盟及其成员国创造一个“繁荣的数据驱动经济”。该决议预计，到 2020 年，欧盟国内生产总值将因更好的数据使用而增加 1.9%。但遗憾的是，据统计目前只有 1.7%的公司充分利用了先进的数字技术。

拓宽和深入大数据技术应用是各国数据战略的共识之处。据了解，美国 2020 年人口普查有望采用差分隐私等大数据隐私保护技术来提高对个人信息的保护。英国政府统计部门正在探索利用交通数据，通过大数据分析及时跟踪英国经济走势，提供预警服务，帮助政府进行精准决策。

（二）大数据底层技术逐步成熟

近年来，大数据底层技术发展呈现出逐步成熟的态势。在大数据发展的初期，技术方案主要聚焦于解决数据“大”的问题，Apache Hadoop 定义了最基础的分布式批处理架构，打破了传统数据库一体

[2] 可参考 https://www.secrss.com/articles/11352

化的模式，将计算与存储分离，聚焦于解决海量数据的低成本存储与规模化处理。Hadoop 凭借其友好的技术生态和扩展性优势，一度对传统大规模并行处理（massively parallel processor, MPP）数据库的市场造成影响。但当前 MPP 在扩展性方面不断突破（2019 年中国信通院大数据产品能力评测中 MPP 大规模测试集群规模已突破 512 节点），使得 MPP 在海量数据处理领域又重新获得了一席之位。

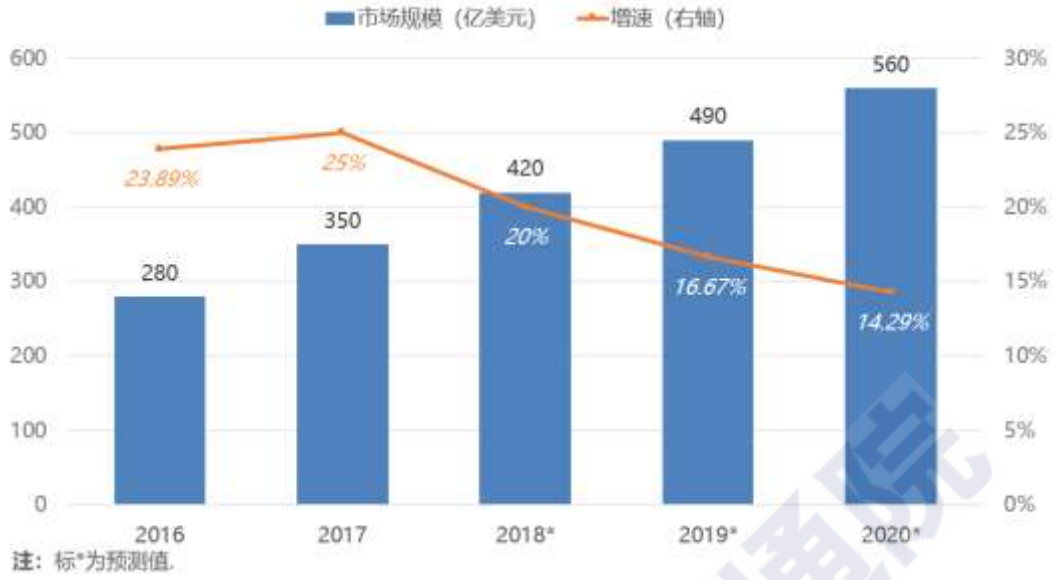
MapReduce 暴露的处理效率问题以及 Hadoop 体系庞大复杂的运维操作，推动计算框架不断进行着升级演进。随后出现的 Apache Spark 已逐步成为计算框架的事实标准。在解决了数据“大”的问题后，数据分析时效性的需求愈发突出，Apache Flink、Kafka Streams、Spark Structured Streaming 等近年来备受关注的产品为流处理的基础框架打下了基础。

在此基础上，大数据技术产品不断分层细化，在开源社区形成了丰富的技术栈，覆盖存储、计算、分析、集成、管理、运维等各个方面。据统计，目前大数据相关开源项目已达上百个。

（三）大数据产业规模平稳增长

国际权威机构 Statista 在 2019 年 8 月发布的报告显示，预计到 2020 年，全球大数据市场的收入规模将达到 560 亿美元，较 2018 年的预期水平增长约 33.33%，较 2016 年的市场收入规模翻一倍。随着市场整体的日渐成熟和新兴技术的不断融合发展，未来大数据市场将呈现稳步发展的态势，增速维持在 14%左右。在 2018-2020 年的预测期内，大数据市场整体的收入规模将保持每年约 70 亿美元的增长，

复合年均增长率约为 15.33%。



数据来源：Wikibon、SiliconANGLE

图 2　2016-2020 年全球大数据市场收入规模预测

从细分市场来看，大数据硬件、软件和服务的市场规模均保持较稳定的增长，预计到 2020 年，三大细分市场的收入规模将分别达到 150 亿美元（硬件）、200 亿美元（软件）、210 亿美元（服务）。具体来看，2016-2017 年，软件市场规模增速达到了 37.50%，在数值上超过了传统的硬件市场。随着机器学习、高级分析算法等技术的成熟与融合，更多的数据应用和场景正在落地，大数据软件市场将继续高速增长。预计在 2018-2020 年间，每年约有 30 亿美元的增长规模，复合年均增长率约为 19.52%。大数据相关服务的规模始终最高，预计在 2018-2020 年间的复合年均增长率约为 14.56%。相比之下，硬件市场增速最低，但仍能保持约 11.80%的复合年均增长率。从整体占比来看，软件规模占比将逐渐增加，服务相关收益将保持平稳发展的趋势，软件与服务之间的差距将不断缩小，而硬件规模在整体的占比则逐渐减小。

数据来源：Wikibon、SiliconANGLE

图 3　2016-2020 年全球大数据细分市场收入规模预测

（四）大数据企业加速整合

近两年来，国际具有影响力的大数据公司也遭遇了一些变化。

2018 年 10 月，美国大数据技术巨头 Cloudera[3]和 Hortonworks[4]宣布合并。在 Hadoop 领域，两家公司的合并意味着“强强联手”，而在更加广义的大数据领域，则更像是“抱团取暖”。但毫无疑问，这至少可以帮助两家企业结束近十年的竞争，并且依靠垄断地位早日摆脱长期亏损的窘状。而从第三方的角度来看，这无疑会影响整个 Hadoop 的生态。开源大数据目前已经成为互联网企业的基础设施，两家公司合并，意味着 Hadoop 的标准将更加统一，长期来看新公司的盈利能力也将大幅提升，并将更多的资源用于新技术的投入。从体量和级别上来看，新公司将基本代表 Hadoop 社区，其他同类型企业将很难与

[3] Cloudera 成立于 2008 年，发行了 Hadoop 集成版本 CDH。CDH 产品包括企业版和开源版，在企业版中，包含闭源管理组件 Cloudera Manager。

[4] Hortonworks 是从雅虎 Hadoop 团队剥离成立的创业公司，不同于 Cloudera，Hortonworks 的软件是完全开源的，通过技术支持来盈利。

之竞争。

2019 年 8 月，惠普（HPE）收购著名大数据技术公司 MapR 的业务资产，包括 MapR 的技术、知识产权以及多个领域的业务资源等。MapR 创立于 2009 年，属于 Hadoop 全球软件发行版供应商之一。专家普遍认为，企业组织越来越多以云服务形式使用数据计算和分析产品是使得 MapR 需求减少的重要原因之一。用户需求正从采购以 Hadoop 为代表的平台型产品，转向结合云化、智能计算后的服务型产品。这也意味着，全球企业级 IT 厂商的战争已经进入到了一个新阶段，即满足用户从平台产品到云化服务，再到智能解决方案的整体需求。

（五）数据合规要求日益严格

近两年来，各国在数据合规性方面的重视程度越来越高，但数据合规的进程仍任重道远。2019 年 5 月 25 日，旨在保护欧盟公民的个人数据、对企业的数据处理提出了严格要求的《通用数据保护条例》（GDPR）实施满一周年，数据保护相关的案例与公开事件数量攀升，同时也引起了诸多争议。

牛津大学的一项研究发现，GDPR 实施满一年后，未经用户同意而设置的新闻网站上的 Cookies 数量下降了 22％[5]。欧盟 EDPB 的报告显示，GDPR 实施一年以来，欧盟当局收到了约 145000 份数据安全相关的投诉和问题举报；共判处 5500 万欧元行政罚款。苹果、微软、Twitter、WhatsApp、Instagram 等企业也都遭到调查或处罚。

[5] 可参考 https://www.cnbeta.com/articles/tech/759403.htm

GDPR 的正式实施之后，带来了全球隐私保护立法的热潮，并成功提升了社会各领域对于数据保护的重视。例如，2020 年 1 月起，美国加州的消费者隐私法案（CCPA）也将正式生效[6]。与 GDPR 类似，CCPA 将对所有和美国加州居民有业务的数据商业行为进行监管。CCPA 在适用监管的标准上比 GDPR 更宽松，但是一旦满足被监管的标准，违法企业受到的惩罚更大。2019 年 8 月份，IAPP/OneTrust 对部分美国企业进行了 CCPA 准备度调查，结果显示，74％的受访者认为他们的企业应该遵守 CCPA，但只有大约 2％的受访者认为他们的企业已经完全做好了应对 CCPA 的准备。除加州 CCPA 外，更多的法案正在美国纽约州等多个州陆续生效。

[6] 可参考 https://www.secrss.com/articles/13773

二、融合成为大数据技术发展的重要特征

当前，大数据体系的底层技术框架已基本成熟。大数据技术正逐步成为支撑型的基础设施，其发展方向也开始向提升效率转变，逐步向个性化的上层应用聚焦，技术的融合趋势愈发明显。本章将针对当前大数据技术的几大融合趋势进行探讨。

（一）算力融合：多样性算力提升整体效率

随着大数据应用的逐步深入，场景愈发丰富，数据平台开始承载人工智能、物联网、视频转码、复杂分析、高性能计算等多样性的任务负载。同时，数据复杂度不断提升，以高维矩阵运算为代表的新型计算范式具有粒度更细、并行更强、高内存占用、高带宽需求、低延迟高实时性等特点，以 CPU 为底层硬件的传统大数据技术无法有效满足新业务需求，出现性能瓶颈。

当前，以 CPU 为调度核心，协同 GPU、FPGA、ASIC 及各类用于 AI 加速“xPU”的异构算力平台成为行业热点解决方案，以 GPU 为代表的计算加速单元能够极大提升新业务计算效率。今年 9 月，腾讯云发布了两款异构计算产品，包括搭载 Xilinx 数据中心加速卡 Alveo U200 的 FPGA 实例 FX4，以及采用 NVIDIA T4 的 GPU 实例 GN7。华为公司计划在今年年底推出 Fusion Insight B160 数据智能模型发布一体化解决方案，内置 Kunpeng920+Atlas300C 芯片，为 AI 模型发布提供强劲算力。

不同硬件体系融合存在开发工具相互独立、编程语言及接口体系不同、软硬件协同缺失等工程问题。为此，产业界试图从统一软件开

发平台和开发工具的层面来实现对不同硬件底层的兼容，例如 Intel 公司正在设计支持跨多架构(包括 CPU、GPU、FPGA 和其他加速器）开发的编程模型 oneAPI，它提供一套统一的编程语言和开发工具集，来实现对多样性算力的调用，从根本上简化开发模式，针对异构计算形成一套全新的开放标准。

（二）流批融合：平衡计算性价比的最优解

流处理能够有效处理即时变化的信息，从而反映出信息热点的实时动态变化。而离线批处理则更能够体现历史数据的累加反馈。考虑到对于实时计算需求和计算资源之间的平衡，业界很早就有了 lambda 架构的理论来支撑批处理和流处理共同存在的计算场景。随着技术架构的演进，流批融合计算正在成为趋势，并不断在向更实时更高效的计算推进，以支撑更丰富的大数据处理需求。

流计算的产生来源于对数据加工时效性的严苛要求。数据的价值随时间流逝而降低时，我们就必须在数据产生后尽可能快的对其进行处理，比如实时监控、风控预警等。早期流计算开源框架的典型工具是 Storm，虽然它是逐条处理的典型流计算模式，但并不能满足“有且仅有一次（Exactly-once）”的处理机制。之后的 Heron 在 Storm 上做了很多改进，但相应的社区并不活跃。同期的 Spark 在流计算方面先后推出了 Spark Streaming 和 Structured Streaming，以微批处理的思想实现流式计算。而近年来出现的 Apache Flink，则使用了流处理的思想来实现批处理，很好地实现了流批融合的计算，国内包括阿里、腾讯、百度、字节跳动，国外包括 Uber、Lyft、Netflix 等公司都是

Flink 的使用者。2017 年由伯克利大学 AMPLab 开源的 Ray 框架也有相类似的思想，由一套引擎来融合多种计算模式，蚂蚁金服基于此框架正在进行金融级在线机器学习的实践。

（三）TA 融合：混合事务/分析支撑即时决策

TA 融合是指事务（Transaction）与分析（Analysis）的融合机制。传统的业务应用在做技术选型时，会根据使用场景的不同选择对应的数据库技术，当应用需要对高并发的用户操作做快速响应时，一般会选择面向事务的 OLTP 数据库；当应用需要对大量数据进行多维分析时，一般会选择面向分析的 OLAP 数据库。

在数据驱动精细化运营的今天，海量实时的数据分析需求无法避免。分析和业务是强关联的，但由于这两类数据库在数据模型、行列存储模式和响应效率等方面的区别，通常会造成数据的重复存储。事务系统中的业务数据库只能通过定时任务同步导入分析系统，这导致了数据时效性不足，无法实时地进行决策分析。

混合事务/分析处理（HTAP）是 Gartner 提出的一个架构，它的设计理念是为了打破事务和分析之间的那堵“墙”，实现在单一的数据源上不加区分的处理事务和分析任务。这种融合的架构具有明显的优势，可以避免频繁的数据搬运操作给系统带来的额外负担，减少数据重复存储带来的成本，从而及时高效地对最新业务操作产生的数据进行分析。

现阶段主流的实现方案主要有三种：一是基于传统的行存关系型数据库（类似 MySQL）实现事务特性，并在此基础上通过引入计算

引擎来增加复杂查询的能力；二是在行存数据库（如 Postgres-XC 版本）的基础上增加列存的功能，来实现分析类业务的需求；三是基于列存为主的分析型数据库（如 Greenplum），增加行存等功能优化，提供事务的支持。但由于没有从根本上改变数据的存储模式，三种方案都会在事务或分析功能上有所侧重，无法完美的在一套系统里互不干扰地处理事务和分析型任务，无法避免对数据的转换和复制，只能在一定程度上缩短分析型业务的时延。

（四）模块融合：一站式数据能力复用平台

大数据的工具和技术栈已经相对成熟，大公司在实战经验中围绕工具与数据的生产链条、数据的管理和应用等逐渐形成了能力集合，并通过这一概念来统一数据资产的视图和标准，提供通用数据的加工、管理和分析能力。

数据能力集成的趋势打破了原有企业内的复杂数据结构，使数据和业务更贴近，并能更快地使用数据驱动决策。主要针对性地解决三个问题：一是提高数据获取的效率；二是打通数据共享的通道；三是提供统一的数据开发能力。这样的“企业级数据能力复用平台”是一个由多种工具和能力组合而成的数据应用引擎、数据价值化的加工厂，来连接下层的数据和上层的数据应用团队，从而形成敏捷的数据驱动精细化运营的模式。阿里巴巴提出的“中台”概念和华为公司提出的“数据基础设施”概念都是模块融合趋势的印证。

（五）云数融合：云化趋势降低技术使用门槛

大数据基础设施向云上迁移是一个重要的趋势。各大云厂商均开始提供各类大数据产品以满足用户需求，纷纷构建自己的云上数据产品。比如 Amazon Web Service(AWS)和 Google Cloud Platform(GCP)很早就开始提供受管理的 MapReduce 或 Spark 服务，以及国内阿里云的 MaxCompute、腾讯云的弹性 MapReduce 等，大规模可扩展的数据库服务也纷纷上云，比如 Google Big Query、AWS Redshift、阿里云的 PolarDB、腾讯云的 Sparkling 等，来为 PB 级的数据集提供分布式数据库服务。早期的云化产品大部分是对已有大数据产品的云化改造，现在，越来越多的大数据产品从设计之初就遵循了云原生的概念进行开发，生于云长于云，更适合云上生态。

向云化解决方案演进的最大优点是用户不用再操心如何维护底层的硬件和网络，能够更专注于数据和业务逻辑，在很大程度上降低了大数据技术的学习成本和使用门槛。

（六）数智融合：数据与智能多方位深度整合

大数据与人工智能的融合则成为大数据领域当前最受关注的趋势之一。这种融合主要体现在大数据平台的智能化与数据治理的智能化。

智能的平台：用智能化的手段来分析数据是释放数据价值高阶之路，但用户往往不希望在两个平台间不断的搬运数据，这促成了大数据平台和机器学习平台深度整合的趋势，大数据平台在支持机器学习算法之外，还将支持更多的 AI 类应用。Databricks 为数据科学家提供一站式的分析平台 Data Science Workspace，Cloudera 也推出了相应

的分析平台 Cloudera Data Science Workbench。2019 年底，阿里巴巴基于 Flink 开源了机器学习算法平台 Alink，并已在阿里巴巴搜索、推荐、广告等核心实时在线业务中有广泛实践。

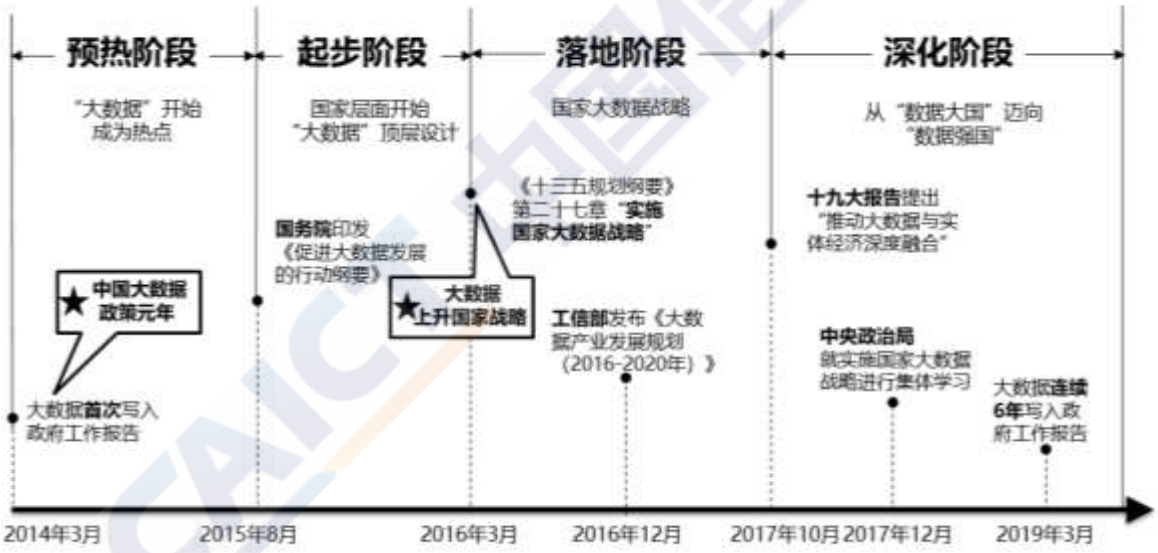
智能的数据治理：数据治理的输出是人工智能的输入，即经过治理后的大数据。数据治理与人工智能的发展存在相辅相成的关系：一方面，数据治理为人工智能的应用提供高质量的合规数据；另一方面，人工智能对数据治理存在诸多优化作用。AI 使能数据治理，是通过智能化的数据治理使数据变得智能：通过智能元数据感知和敏感数据自动识别，对数据自动分级分类，形成全局统一的数据视图。通过智能化的数据清洗和关联分析，把关数据质量，建立数据血缘关系。数据能够自动具备类型、级别、血缘等标签，在降低数据治理复杂性和成本的同时，得到智能的数据。

三、大数据产业蓬勃发展

近年来，我国大数据产业蓬勃发展，融合应用不断深化，数字经济量质提升，对经济社会的创新驱动、融合带动作用显著增强。本章将从政策环境、主管机构、产品生态、行业应用等方面对我国大数据产业发展的态势进行简要分析。

（一）大数据产业发展政策环境日益完善

产业发展离不开政策支撑。我国政府高度重视大数据的发展。自2014年以来，我国国家大数据战略的谋篇布局经历了四个不同阶段。



来源：中国信息通信研究院

图 4　国家大数据战略的布局历程

- 预热阶段：2014 年 3 月，“大数据”一词首次写入政府工作报告，为我国大数据发展的政策环境搭建开始预热。从这一年起，“大数据”逐渐成为各级政府和社会各界的关注热点，中央政府开始提供积极的支持政策与适度宽松的发展环境，为大数据发展创造机遇。

- 起步阶段：2015 年 8 月 31 日，国务院正式印发了《促进大数据发展行动纲要》（国发〔2015〕50 号），成为我国发展大数据的首部战略性指导文件，对包括大数据产业在内的大数据整体发展作出了部署，体现出国家层面对大数据发展的顶层设计和统筹布局。
- 落地阶段：《十三五规划纲要》的公布标志着国家大数据战略的正式提出，彰显了中央对于大数据战略的重视。2016 年 12 月，工信部发布《大数据产业发展规划（2016-2020 年）》，为大数据产业发展奠定了重要的基础。
- 深化阶段：随着国内大数据迎来全面良好的发展态势，国家大数据战略也开始走向深化阶段。2017 年 10 月，党的十九大报告中提出推动大数据与实体经济深度融合，为大数据产业的未来发展指明方向。12 月，中央政治局就实施国家大数据战略进行了集体学习。2019 年 3 月，政府工作报告第六次提到“大数据”，并且有多项任务与大数据密切相关。

自 2015 年国务院发布《促进大数据发展行动纲要》系统性部署大数据发展工作以来，各地陆续出台促进大数据产业发展的规划、行动计划和指导意见等文件。截至目前，除港澳台外全国 31 个省级单位均已发布了推进大数据产业发展的相关文件。可以说，我国各地推进大数据产业发展的设计已经基本完成，陆续进入了落实阶段。以下我们将 31 个省级行政单位的典型大数据产业政策进行总结。

表 1　全国 31 省级行政单位代表性大数据产业政策

省级单位	政策	发布时间
北京	北京市大数据和云计算发展行动计划	2016 年 8 月 3 日
上海	上海市大数据发展实施意见	2016 年 9 月 15 日
天津	天津市促进大数据发展应用条例	2018 年 12 月 14 日
重庆	重庆市以大数据智能化为引领的创新驱动发展战略行动计划（2018-2020 年）	2018 年 8 月 23 日
广东	广东省促进大数据发展行动计划(2016-2020 年)	2016 年 4 月 22 日
福建	福建省促进大数据发展实施方案(2016-2020 年)	2016 年 6 月 18 日
浙江	浙江省促进大数据发展实施计划	2016 年 2 月 18 日
江苏	江苏省大数据发展行动计划	2016 年 8 月 19 日
山东	关于促进大数据发展的实施意见	2017 年 5 月 23 日
河北	河北省大数据产业创新发展三年行动计划（2018-2020 年）	2018 年 3 月 22 日
辽宁	辽宁省运用大数据加强对市场主体服务和监管实施方案	2015 年 10 月 19 日
吉林	关于运用大数据加强对市场主体服务和监管的实施意见	2016 年 5 月 25 日
黑龙江	黑龙江省促进大数据发展三年行动计划	2017 年 12 月 11 日
内蒙古	内蒙古自治区大数据发展总体规划（2017-2020 年）	2017 年 12 月 28 日
甘肃	甘肃省数据信息产业发展专项行动计划	2018 年 6 月 3 日
新疆	新疆维吾尔自治区云计算与大数据产业“十三五”发展规划	2016 年 12 月 8 日
云南	关于重点行业和领域大数据开放开发工作的指导意见	2017 年 6 月 23 日
广西	促进大数据发展行动方案	2017 年 5 月 22 日
贵州	关于促进大数据云计算人工智能创新发展加快建设数字贵州的意见	2018 年 6 月 21 日
四川	四川省促进大数据发展工作方案	2018 年 1 月 4 日
青海	关于印发促进云计算发展培育大数据产业实施意见的通知	2015 年 8 月 10 日
宁夏	宁夏回族自治区大数据产业发展条例(征求意见稿)	2017 年 5 月 5 日
山西	山西省大数据发展规划（2017-2020 年）	2017 年 3 月 13 日
河南	河南省大数据产业发展三年行动计划（2018-2020 年）	2018 年 5 月 9 日
安徽	安徽省运用大数据加强对市场主体服务和监管实施方案	2015 年 10 月 30 日
江西	江西省大数据发展行动计划	2017 年 7 月 5 日
湖南	湖南省大数据产业发展三年行动计划（2019-2021 年）	2019 年 1 月 24 日
湖北	湖北省大数据发展行动计划（2016-2020 年）	2016 年 9 月 14 日

陕西	大数据与云计算产业示范工程实施方案	2016 年 6 月 17 日
海南	海南省促进大数据发展实施方案	2016 年 11 月 25 日
西藏	西藏自治区人民政府关于推动云计算应用大数据发展培育经济发展新动力的意见	2017 年 7 月 10 日

来源：中国信息通信研究院

需要说明的是，大部分省（区、市）都发布了不止一项大数据相关政策，以上所列的只是其中最主要的一项。可以看出，大部分省（区、市）的大数据政策集中发布于 2016 年至 2017 年。而在近两年发布的政策中，更多的地方将新一代信息技术整体作为考量，并加入了人工智能、数字经济等内容，进一步地拓展了大数据的外延。同时，各地在颁布大数据政策时，除注重大数据产业的推进外，也在更多地关注产业数字化和政务服务等方面，这也体现出了大数据与行业应用结合及政务数据共享开放近年来取得的进展。

（二）各地大数据主管机构陆续成立

近年来，部分省市陆续成立了大数据局等相关机构，对包括大数据产业在内的大数据发展进行统一管理。以省级大数据主管机构为例，从 2014 年广东省设立第一个省级大数据局开始，截至 2019 年 5 月，共有 14 个省级地方成立了专门的大数据主管机构。

省级大数据主管机构的设立过程可以分为两个阶段。第一个阶段从 2014 年 2 月至 2018 年上半年。2014 年 2 月，广东省在全国率先成立了广东省大数据管理局，成为第一个省级大数据管理局。2015 年，贵州省和浙江省先后成立了贵州省大数据发展管理局和浙江省数据管理中心。其中，贵州省大数据发展管理局是首个省政府直属的大数据治理机构。2017 年，省级大数据治理机构又增加了 4 个，分别

是内蒙古自治区大数据发展管理局、重庆市大数据发展局、江西省大数据中心、陕西省政务数据服务局。2018 年 6 月，上海、天津两个直辖市分别成立了上海市大数据中心和天津市大数据管理中心。

第二阶段开始于 2018 年下半年。按照中央部署，新一轮省级机构改革方案陆续发布，各地纷纷以不同的方式组建或调整政府数据治理机构。其中，一部分省(市、自治区)陆续成立了专门的大数据管理机构。另一部分省(市、自治区)则是对原有机构进行了调整组合。

表 2 省级大数据主管机构[7]

行政区	设立时间	机构名称	隶属机构	机构性质
广东	2018 年	广东省政务服务数据管理局[8]	广东省人民政府办公厅	政府部门的管理机构
贵州	2015 年	贵州省大数据发展管理局	贵州省人民政府	政府直属机构
浙江	2018 年	浙江省大数据发展管理局	浙江省人民政府办公厅	政府部门的管理机构
内蒙古	2017 年	内蒙古自治区大数据发展管理局	内蒙古自治区人民政府	政府直属机构
重庆	2018 年	重庆市大数据应用发展管理局[9]	重庆市人民政府	政府直属机构
陕西	2017 年	陕西省政务数据服务局[10]	陕西省人民政府	政府直属机构
福建	2018 年	数字福建建设领导小组办公室（福建省大数据管理局）	福建省发展和改革委员会	政府部门的管理机构
广西	2018 年	广西壮族自治区大数据发展局	广西壮族自治区人民政府	政府直属机构
山东	2018 年	山东省大数据局	山东省人民政府	政府直属机构
北京	2018 年	北京市经济和信息化局（北京市大数据管理局）	北京市人民政府	政府组成部门
安徽	2018 年	安徽省数据资源管理局（安徽省政务服务管理局）	安徽省人民政府	政府直属机构
河南	2018 年	河南省大数据管理局	河南省人民政府办公厅	政府部门的管理机构

[7] 黄璜，孙学智.中国地方政府数据治理机构的初步研究：现状与模式[J].中国行政管理，2018（12）：31-36.
[8] 广东省最初于 2014 年设立了广东省大数据管理局，隶属于广东省经济和信息化委员会。
[9] 重庆市最初于 2017 年设立了重庆市大数据发展管理局，隶属于重庆市经济和信息化委员会。
[10] 陕西省工业和信息化厅加挂陕西省政务数据服务局牌子，相关职能主要由陕西省大数据管理与服务中心承担。

吉林	2018 年	吉林省政务服务和数字化建设管理局	吉林省人民政府	政府直属机构
海南	2019 年	海南省大数据管理局	海南省人民政府	政府组成部门[11]

来源：中国信息通信研究院

除此之外，上海、天津、江西等省市组建了上海市大数据中心、天津市大数据管理中心、江西省信息中心（江西省大数据中心），承担了一部分大数据主管机构的职能。

部分省级以下的地方政府也相应组建了专门的大数据管理机构。根据黄璜等人的统计[12]，截至 2018 年 10 月已有 79 个副省级和地级城市组建了专门的大数据管理机构。

根据机构隶属关系，地方政府大数据主管机构可以大致分为三类。一是作为政府组成部门。例如，北京市大数据管理局由北京市经济和信息化局加挂牌子，隶属于北京市人民政府，是政府的组成部门。这种情况下，大数据局的行政职能相对较强，级别和权责水平也相对较高。二是作为政府直属机构。例如，内蒙古自治区大数据发展管理局虽隶属于自治区人民政府，但其作为政府的直属机构，更多承担事业单位的相关职能。三是作为政府部门的管理机构。例如，广东省政务服务数据管理局隶属于广东省人民政府办公厅，是政府部门的下属机构。

根据组建模式，地方政府大数据主管机构可以大致分为五类。一是以地方发改委为基础进行组建。这种类型的大数据主管机构较多，其优势在于可以更好地承担地方大数据宏观管理和相关项目审批职

[11] 海南省大数据管理局由海南省政府依法设立，但不列入行政机构序列，不从事法定职责外事务，是具有独立法人地位的法定机构。

[12] 黄璜，孙学智.中国地方政府数据治理机构的初步研究：现状与模式[J].中国行政管理，2018（12）：31-36.

能。二是对政府办公室(厅)相关职能进行重组。这种类型的大数据主管机构的优势在于政府系统信息化建设经验丰富，对于推动电子政务建设优势突出。三是对原有信息中心进行重组。这种类型的大数据主管机构的优势在于直接接触数据资源较多，便于开展区域内大数据资源的统筹管理工作。四是以地方经信委/工信厅为基础进行组建。这种类型的大数据主管机构在推动大数据产业发展方面具有得天独厚的优势。五是对原有机构增加相关职能，即原有机构基础上加挂牌子，但可能会专门设立几个承担大数据管理职能的处室。这种类型的大数据主管机构其核心职能仍然是原有机构的主要职能，便于与原有工作的衔接。

由于地方大数据主管机构在隶属机构和组建模式上的不同，其机构职责也不尽相同。大多数机构都包含制订地方大数据战略规划的职能，但在产业发展政策制订、数据资源整合、数据资源开放共享、电子政务系统建设、信息安全、政府网站建设等方面的职能则并非所有大数据主管机构都具备。

（三）大数据技术产品水平持续提升

从产品角度来看，目前大数据技术产品主要包括**大数据基础类技术产品**（承担数据存储和基本处理功能，包括分布式批处理平台、分布式流处理平台、分布式数据库、数据集成工具等）、**分析类技术产品**（承担对于数据的分析挖掘功能，包括数据挖掘工具、BI 工具、可视化工具等）、**管理类技术产品**（承担数据在集成、加工、流转过程中的管理功能，包括数据管理平台、数据流通平台等）等。我国在

这些方面都取得了一定的进展。

我国大数据基础类技术产品市场成熟度相对较高。一是供应商越来越多，从最早只有几家大型互联网公司发展到目前的近 60 家公司可以提供相应产品，覆盖了互联网、金融、电信、电力、铁路、石化、军工等不同行业；二是产品功能日益完善，根据中国信通院的测试，分布式批处理平台、分布式流处理平台类的参评产品功能项通过率均在 95%以上；三是大规模部署能力有很大突破，例如阿里云 MaxCompute 通过了 10000 节点批处理平台基础能力测试，华为 GuassDB 通过了 512 台物理节点的分析型数据库基础能力测试；四是自主研发意识不断提高，目前有很多基础类产品源自对于开源产品进行的二次开发，特别是分布式批处理平台、流处理平台等产品九成以上基于已有开源产品开发。

我国大数据分析类技术产品发展迅速，个性化与实用性趋势明显。一是满足跨行业需求的通用数据分析工具类产品逐渐应运而生，如百度的机器学习平台 Jarvis、阿里云的机器学习平台 PAI 等；二是随着深度学习技术的相应发展，数据挖掘平台从以往只支持传统机器学习算法转变为额外支持深度学习算法以及 GPU 计算加速能力；三是数据分析类产品易用性进一步提升，大部分产品都拥有直观的可视化界面以及简洁便利的交互操作方式。

我国大数据管理类技术产品还处于市场形成的初期。目前，国内常见的大数据管理类软件有 20 多款。数据管理类产品虽然涉及的内容庞杂，但技术实现难度相对较低，一些开源软件如 Kettle、Sqoop

和 Nifi 等，为数据集成工具提供了开发基础。中国信通院测试结果显示，参照囊括功能全集的大数据管理软件评测标准，所有参评产品符合程度均在 90%以下。随着数据资产的重要性日益突出，数据管理类软件的地位也将越来越重要，未来将机器学习、区块链等新技术与数据管理需求结合，还有很大的发展空间。

（四）大数据行业应用不断深化

前几年，大数据的应用还主要在互联网、营销、广告领域。而随着大数据工具的门槛降低以及企业数据意识的不断提升，越来越多的行业开始尝到大数据带来的“甜头”。这几年，无论是从新增企业数量、融资规模还是应用热度来说，与大数据结合紧密的行业逐步向工业、政务、电信、交通、金融、医疗、教育等领域广泛渗透，应用逐渐向生产、物流、供应链等核心业务延伸，涌现了一批大数据典型应用，企业应用大数据的能力逐渐增强。电力、铁路、石化等实体经济领域龙头企业不断完善自身大数据平台建设，持续加强数据治理，构建起以数据为核心驱动力的创新能力，行业应用“脱虚向实”趋势明显，大数据与实体经济深度融合不断加深。

电信行业方面，电信运营商拥有丰富的数据资源，除了传统经营模式下存在于 BOSS、CRM 等经营系统的结构化数据，还包括移动互联网业务经营形成的文本、图片、音视频等非结构化数据。数据来源涉及移动通话和固定电话、无线上网、有线宽带接入等所有业务，也涵盖线上线下渠道在内的渠道经营相关信息，所服务的客户涉及个人客户、家庭客户和政企客户。三大运营商 2019 年以来在大数据应

用方面都走向了更加专业化的阶段。电信行业在发展大数据上有明显的优势，主要体现数据规模大、数据应用价值持续凸显、数据安全性普遍较高。2019 年，三大运营商都已经完成了全集团大数据平台的建设，设立了专业的大数据运营部门或公司，开始了数据价值释放的新举措。通过对外提供领先的网络服务能力，深厚的数据平台架构和数据融合应用能力，高效可靠的云计算基础设施和云服务能力，打造数字生态体系，加速非电信业务的变现能力。

金融行业方面，随着金融监管日趋严格，通过金融大数据规范行业秩序并降低金融风险逐渐成为金融大数据的主流应用场景。同时，各大金融机构由于信息化建设基础好、数据治理起步早，使得金融业成为数据治理发展较为成熟的行业。

互联网营销方面，随着社交网络用户数量的不断扩张，利用社交大数据来做产品口碑分析、用户意见收集分析、品牌营销、市场推广等“数字营销”应用，将会是未来大数据应用的重点。电商数据直接反映用户的消费习惯，具有很高的应用价值。伴随着移动互联网流量见顶，以及广告主营销预算的下降，如何利用大数据技术帮助企业更高效地触达目标用户成为行业最为热衷的话题。“线下大数据”“新零售”的概念日渐火热。但其对于个人信息保护方面容易存在漏洞，也使得合规性成为这一行业发展的核心问题。

工业方面，工业大数据是指在工业领域里，在生产链过程包括研发、设计、生产、销售、运输、售后等各个环节中产生的数据总和。工业大数据来源主要有三类，一是生产经营相关数据，主要存储于企

业信息系统内部，涵盖传统工业设计和制造类软件、客户关系管理（CRM）、供应链管理（SCM）、产品生命周期管理（PLM）等；二是设备物联数据，主要包括物联网运行模式下工业生产设备和目标产品实时运行数据、设备和产品运行状态相关数据；三是外部相关数据，主要涵盖与工业主体生产活动和产品相关的企业外部数据。设备故障预测、能耗管理、智能排产、库存管理和供应链协同一直是工业大数据应用的主攻方向。随着工业大数据成熟度的提升，工业大数据的价值挖掘也逐渐深入。目前，各个工业企业已经开始面向数据全生命周期的数据资产管理，逐步提升工业大数据成熟度，深入工业大数据价值挖掘。

能源行业方面，2019 年 5 月，国家电网大数据中心正式成立，该中心旨在打通数据壁垒、激活数据价值、发展数字经济，实现数据资产的统一运营，推进数据资源的高效使用。这是传统能源行业拥抱大数据应用的一次机制创新。

医疗健康方面，医疗大数据成为 2019 年大数据应用的热点方向。2018 年 7 月颁布的《国家健康医疗大数据标准、安全和服务管理办法》为健康行业大数据服务指导了方向。电子病历、个性化诊疗、医疗知识图谱、临床决策支持系统、药品器械研发等成为行业热点。

除以上行业之外，教育、文化、旅游等各行各业的大数据应用也都在快速发展。我国大数据的行业应用更加广泛，正加速渗透到经济社会的方方面面。

四、数据资产化步伐稳步推进

在中国共产党十九届四中全会上，中央首次公开提出“健全劳动、资本、土地、知识、技术、管理和数据等生产要素按贡献参与分配的机制。”这是中央首次在公开场合提出数据可作为生产要素按贡献参与分配，反映了随着经济活动数字化转型加快，数据对提高生产效率的乘数作用凸显，成为最具时代特征新生产要素的重要变化[13]。

（一）数据：从资源到资产

“数据资产”这一概念是由信息资源和数据资源的概念逐渐演变而来的[14]。信息资源是在 20 世纪 70 年代计算机科学快速发展的背景下产生的，信息被视为与人力资源、物质资源、财务资源和自然资源同等重要的资源，高效、经济地管理组织中的信息资源是非常必要的。数据资源的概念是在 20 世纪 90 年代伴随着政府和企业的数字化转型而产生，是有含义的数据集结到一定规模后形成的资源。数据资产在 21 世纪初大数据技术的兴起背景下产生，并随着数据管理、数据应用和数字经济的发展而普及。

中国信通院在 2017 年发布了《数据资产管理实践白皮书》，其中将“数据资产”定义为“由企业拥有或者控制的，能够为企业带来未来经济利益的，以一定方式记录的数据资源”。这一概念强调了数据具备的“预期给会计主体带来经济利益”的资产特征。

企业中数据资产的概念边界随着数据管理技术的变化而不断拓

[13] 刘鹤，《坚持和完善社会主义基本经济制度》，2019.11。
[14] 叶雅珍，刘国华，朱扬勇，《数据资产相关概念综述》，2019.11。

展。在早期文件系统阶段，数据以“文件”的形式保存在磁盘之上，初步实现了数据的访问和长期保存，数据资产主要指这些存储的“文件”。随后的数据库与数据仓库阶段，数据资产主要指结构化数据，包括业务数据和各类分析报表等，被用来支撑企业的经营和高层各项决策。在大数据阶段，随着分布式存储、分布式计算以及多种 AI 技术的应用，结构化数据之外的数据也被纳入到数据资产的范畴，数据资产边界拓展到了海量的标签库、企业级知识图谱、文档、图片、视频等内容。

（二）数据资产管理理论体系仍在发展

数据管理的概念是伴随着上世纪八十年代数据随机存储技术和数据库技术的使用而诞生的，主要指在计算机系统中的数据可以被方便地存储和访问。经过 40 年的发展，数据管理的理论体系主要形成了国际数据管理协会（DAMA）、IBM 和数据管控机构（DGI）所提出的三个流派，如下表所示。

表 3　数据管理框架对比

框架名称	内涵	主要观点
DAMA 数据管理框架	是指规划、控制和提供数据资产的一组业务职能，包括开发、执行和监督有关数据的计划、政策、方案、项目、流程、方法和程序，从而控制、保护、交付和提高数据资产的价值	管理活动是组织执行的任务，环境要素是管理活动能够正常执行的基础，二者的匹配性决定了数据管理能否成功。 框架更偏向知识梳理，覆盖比较全面，但缺少可执行的步骤。
IBM 数据治理模型	是指组织管理其信息知识并回答问题的能力	从可执行的角度梳理了 4 个功能域，明确了数据管理的最终目的，具有一定的可执行性。
DGI 数据治理框架	是包含信息相关过程的决策权和控制的活动集合	建立标准的流程，协调降低成本，确保流程的透明性。 框架较为清晰，确立了目的、活动主体和行为准则。

来源：中国信息通信研究院

然而，以上三种理论体系都是大数据时代之前的产物，其视角还是将数据作为信息来管理，更多的是为了满足监管要求和企业考核的目的，并没有从数据价值释放的维度来考虑。

在数据资产化背景下，数据资产管理是在数据管理基础上的进一步发展，可以视作数据管理的“升级版”。主要区别表现为以下三方面。**一是管理视角不同，**数据管理主要关注的是如何解决问题数据带来的损失，而数据资产管理则关注如何利用数据资产为企业带来价值，需要基于数据资产的成本、收益来开展数据价值管理。**二是管理职能不同，**传统数据管理的管理职能包含数据标准管理、数据质量管理、元数据管理、主数据管理、数据模型管理、数据安全管理等，而数据资产管理针对不同的应用场景和大数据平台建设情况，增加了数据价值管理和数据共享管理等职能。**三是组织架构不同，**在“数据资源管理转向数据资产管理”的理念影响下，相应的组织架构和管理制度也有所变化，需要有更专业的管理队伍和更细致的管理制度来确保数据资产管理的流程性、安全性和有效性。

（三）各行业积极实践数据资产管理

各行业实践数据资产管理普遍经历三至四个阶段。最初，行业数据资产管理主要是为了解决报表和经营分析的准确性，并通过建立数据仓库实现。随后，行业数据资产管理的目的是治理数据，管理对象由分析域延伸到生产域，并在数据库中开展数据标准管理和数据质量管理。随着大数据技术的发展，企业数据逐步汇总到大数据平台，形成了数据采集、计算、加工、分析等配套工具，建立了元数据管理、

数据共享、数据安全保护等机制，并开展了数据创新应用。而目前，许多行业的数据资产管理已经进入到数据资产运营阶段，数据成为了企业核心的生产要素，不仅满足企业内部各项业务创新，还逐渐成为服务企业外部的数据产品。企业也积极开展如数据管理能力成熟度模型（DCMM）等数据管理能力评估工作，不断提升数据资产管理能力。

金融、电信等行业普遍在2000年至2010年间就开始了数仓建设，并将数据治理范围逐步扩展到生产域，建立了比较完善的数据治理体系。2010 年后通过引入大数据平台，企业实现了数据的汇聚，并逐渐向数据湖发展，内部的数据应用较为完善，不少企业逐渐在探索数据对外运营和服务。四大国有银行都单独成立了主管数据的一级部门（管理信息部或数据管理部），负责数据资产管理与应用、监管数据报送和外部数据的合作等工作。其它银行的数据管理工作多数由科技部门负责，部分由业务部门负责。2018 年银保监会发布的《银行业金融机构数据治理指引》，强化了银行业数据治理和数据资产管理的工作力度。三大电信运营商在信息化部下成立了数据中心部门来统一数据能力的建设，近年来，除了满足内部的数据应用外，还积极向外拓展，电信和联通都成立了专业的数据对外服务的公司，通过开放平台和数据产品来服务外部客户。

能源、工业等领域数据治理的起步较晚，大都在 2010 年后开始数仓建设。南方电网经历了早期标准化阶段、一体化应用解决、数据质量提升阶段和数据资产管理体系建设四个阶段，当前已经形成了丰

富的制度体系和技术工具体系。国家电网多年前建立了大数据团队，内部自研大数据产品，建设统一数据应用中心，研发了大量的电力大数据应用，2019 年 5 月成立了专业的大数据中心，围绕泛在电力物联网，建设省级和总部两级联动的数据中台能力，目前也在构建较为完善的数据资产管理体系。而根据中国信通院与工业互联网产业联盟的联合调查[15]，工业企业的数据资产管理实际落地情况不甚乐观，超过半数的工业企业的数据资产管理仍处于起步规划阶段，近半数的工业企业未设置数据资产管理专职机构。

（四）数据资产管理工具百花齐放

数据资产管理工具是数据资产管理工作落地的重要手段。由于大数据技术栈中开源软件的缺失，数据资产管理的技术发展没有可参考的模板，工具开发者多从数据资产管理实践与项目中设计工具架构，各企业数据资产管理需求的差异化使得数据资产管理工具的形态各异。因此，数据资产管理工具市场呈现百花齐放的状态。

数据资产管理工具可以是多个工具的集成，并以模块化的形式集中于数据管理平台。国际机构 Gartner 连续三年分别发布了元数据管理工具、数据质量管理工具、分析和商务智能工具的魔力象限[16]，总结了各类工具的核心功能，从而针对性的解决企业在数据资产管理中的问题。由于任何一类工具只能完成数据资产管理的某一项活动，因此供应商往往以数据资产管理工具组合的方式提供服务。数据管理平

[15] 《2018 工业企业数据资产管理现状调查报告》，工业互联网产业联盟/中国中国信通院，2018 年 12 月。
[16] 指在某一特定时间内的对市场情况进行的图形化描述。

台作为数据资产管理工具的集成平台，将各类工具以模块化的方式嵌入平台，并通过对各模块建立关联，实现了数据的全流程、多维度管理。

元数据管理工具、数据标准管理工具、数据质量管理工具是数据资产管理工具的核心，数据价值工具是数据资产化的有力保障。中国信通院对数据管理平台的测试结果显示，数据管理平台对于元数据管理工具、数据标准管理工具和数据质量管理工具的覆盖率达到了100%，这些工具通过追踪记录数据、标准化数据、稽核数据的关键活动，有效地管理了数据，提升了数据的可用性。与此同时，主数据管理工具和数据模型管理工具的覆盖率均低于 20%，其中主数据管理多以解决方案的方式提供服务，而数据模型管理多在元数据管理中实现，或以独立工具在设计数据库或数据仓库阶段完成。超过 80%的数据价值工具以直接提供数据源的方式进行数据服务，其它的数据服务方式包括数据源组合、数据可视化和数据算法模型等。超过 95%的数据价值工具动态展示数据的分布应用和存储计算情况，但仅有不到10%的工具量化数据价值，并提供数据增值方案。

未来，数据资产管理工具将向智能化和敏捷化发展，并以自助服务分析的方式深化数据价值。Gartner 在 2019 年关于分析与商务智能软件市场的调研报告中显示，该市场在 2018 年增长了 11.7%，而基于自助服务分析的现代商务智能和数据科学平台分别增长了 23.3%和 19.0%。随着数据量的增加和数据应用场景的丰富，数据间的关系变得更加复杂，问题数据也隐藏于数据湖中难以被发觉。智能化的探

索梳理结构化数据间、非结构化数据间的关系将节省巨大的人力，快速发现并处理问题数据也将极大的提升数据的可用性。在数据交易市场尚未成熟的情况下，通过扩展数据使用者的范围，提升数据使用者挖掘数据价值的能力，将最大限度地开发和释放数据价值。

国际数据管理协会（DAMA）和中国信通院提出的数据资产管理框架，以及国外先进数据管理工具成为了国内数据资产管理工具研发的主要参考依据。据中国信通院测试和调研统计，国内数据资产管理的商业化产品超过了 30 款，中国信通院的测试覆盖了其中的三分之一。同时，由于金融、电信、互联网等行业具备较强的研发能力和数据资产管理需求，中国信通院的测试也囊括了部分自产自用的数据资产管理产品。

（五）数据资产化面临诸多挑战

目前，困扰数据资产化的关键问题主要包括数据确权困难、数据估值困难和数据交易市场尚未成熟。

1. 数据确权困难

明确数据权属是数据资产化的前提[17]，但目前在数据权利主体以及权力分配上存在诸多争议。

数据权不同于传统物权。物权的重要特征之一是对物的直接支配，但数据权在数据的全生命周期中有不同的支配主体[18]，有的数据产生之初由其提供者支配，有的产生之初便被数据收集人支配（如微信聊

[17] 陈一，《我国大数据交易产权管理实践及政策进展研究》，现代情报，2019。
[18] 可参考 http://www.sohu.com/a/203070639_655070

天内容、电商消费数据、物流数据等）；在数据处理阶段被各类数据主体所支配。原始数据只是大数据产业的基础，其价值属性远低于集合数据为代表的增值数据所产生的价值。

因此，法律专家们倾向于将数据的权属分开，即不探讨整体数据权，而是从管理权、使用权、所有权等维度进行探讨。而由于数据从法律上目前尚没有被赋予资产的属性，所以数据所有权、使用权、管理权、交易权等权益没有被相关的法律充分认同和明确界定。数据也尚未像商标、专利一样，有明确的权利申请途径、权利保护方式等，对于数据的法定权利，尚未有完整的法律保护体系。

2. 数据估值困难

影响数据资产价值的因素主要有质量、应用和风险三个维度[19]。质量是决定数据资产价值的基础，合理评估数据的质量水平，才能对数据的应用价值进行准确预测；应用是数据资产形成价值的方式，数据与应用场景结合才能贡献经济价值；风险则是指法律和道德等方面存在的限制。

表 4　数据价值的影响因素

主要维度	要点
质量维度	·真实性：表示数据的真实程度； ·完整性：表示数据对被记录对象的所有相关指标的完整程度； ·准确性：表示数据被记录的准确性； ·数据成本：在数据交易市场不活跃的情况下，数据价值没有明确的计算方式，卖方出售数据会首先考虑获取数据时的成本； ·安全性：表示数据不被窃取或破坏的能力。
应用维度	·稀缺性：表示数据资产拥有者对数据的独占程度； ·时效性：决定依据数据做出的决策在特定时间内是否有效； ·多维性：表示数据涵盖范围的多样性；

[19] 《数据资产的估值与行业实践》，德勤中国、阿里巴巴集团，2019 年 10 月。

	·场景经济性：指在不同应用场景下，数据所贡献的经济价值也有所不同。
风险维度	·法律限制：在法律尚未明确规定的情况下，哪些数据绝对不能交易，哪些数据可以通过设计合法后才能交易，这些问题在限制数据交易的同时也影响着数据资产的价值； ·道德约束：部分数据交易存在一定程度的道德风险。

来源：中国信息通信研究院

目前，常用的数据资产估值方法主要有成本法、收益法和市场法三类。成本法从资产的重置角度出发，重点考虑资产价值与重新获取或建立该资产所需成本之间的相关程度；收益法基于目标资产的预期应用场景，通过未来产生的经济效益的折现来反映数据资产在投入使用后的收益能力，而根据衡量无形资产经济效益的不同方法又可具体分为权利金节省法、多期超额收益法和增量收益法；市场法则是在相同或相似资产的市场可比案例的交易价格的基础上，对差异因素进行调整，以此反映数据资产的市场价值。

评估数据资产的价值需要考虑多方面因素，数据的质量水平、不同的应用场景和特定的法律道德限制均对数据资产价值有所影响。虽然目前已有从不同角度出发的数据资产估值方法，但在实际应用中均存在不同的问题，有其适用性的限制。构建成熟的数据资产评价体系，还需要以现有方法为基础框架，进一步探索在特定领域和具体案例中的适配方法。

3. 数据交易市场尚未成熟

2014 年以来，国内出现了一批数据交易平台，各地方政府也成立了数据交易机构，包括贵阳大数据交易所、长江大数据交易中心、上海数据交易中心等。同时，互联网领军企业也在积极探索新的数据

流通机制，提供了行业洞察、营销支持、舆情分析、引擎推荐、API 数据市场等数据服务，并针对不同的行业提出了相应的解决方案。

但是，由于数据权属和数据估值的限制，以及数据交易政策和监管的缺失等因素，目前国内的数据交易市场尽管在数据服务方式上有所丰富，却发展依然面临诸多困难，阻碍了数据资产化的进程。主要体现在如下两点。一是市场缺乏信任机制，技术服务方、数据提供商、数据交易中介等可能会私下缓存并对外共享、交易数据，数据使用企业不按协议要求私自留存、复制甚至转卖数据的现象普遍存在。我国各大数据交易平台并未形成统一的交易流程，甚至有些交易平台没有完整的数据交易规范，使得数据交易存在很大风险。二是缺乏良性互动的数据交易生态体系。数据交易中所涉及的采集、传输、汇聚活动日益频繁，相应的，个人隐私、商业机密等一系列安全问题也日益突出，亟需建立包括监管机构和社会组织等多方参与的，法律法规和技术标准多要素协同的，覆盖数据生产流通全过程和数据全生命周期管理的数据交易生态体系。

五、数据安全合规要求不断提升

2019 年以来，大数据安全合规方面不断有事件曝出。9 月 6 日，位于杭州的大数据风控平台杭州魔蝎数据科技有限公司被警方控制，高管被带走，相关服务暂时瘫痪[20]。同日，另一家提供大数据风控服务的新颜科技人工智能科技有限公司高管被带走协助调查。以两平台被查为开端，短短一周内，多家征信企业分别有人被警方带走调查，市场纷纷猜测是否与爬虫业务有关。一时间，大数据安全合规的问题，特别是对于个人信息保护的问题，再次成为了行业关注热点。

（一）数据相关法律监管日趋严格规范

与全球不断收紧的数据合规政策相类似，我国在数据法律监管方面也日趋严格规范。

当前我国大数据方面的立法呈现出以个人信息保护为核心，包含基本法律、司法解释、部门规章、行政法规等综合框架。一些综合性法律中也涉及了个人信息保护条款。

表 5　我国大数据相关立法

法律层级	主要相关法律
基本法律	《中华人民共和国网络安全法》和《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》等
司法解释	主要包括《关于办理侵犯公民个人信息刑事案件适用法律若干问题的解释》、《最高人民法院关于审理利用信息网络侵害人身权益民事纠纷案件适用法律若干问题的规定》等
部门规章	主要包括《电信和互联网用户个人信息保护规定》、《中国人民银行关于银行业金融机构做好个人金融信息保护工作的通知》等
行政法规	主要包括《征信业管理条例》等
综合性法律	在《民法总则》、《刑法修正案（九）》、《侵权责任法》、《消费者权益保护法》、《反恐怖主义法》等综合性法律中，也有涉及个人信息保护的相关条款

来源：中国信息通信研究院

[20] 信息来源于公开网络报道。

2019 年以来，数据安全方面的立法进程明显加快。中央网信办针对四项关于数据安全的管理办法相继发布征求意见稿，其中，《儿童个人信息网络保护规定》已正式公布，并于 10 月 1 日开始施行。一系列行政法规的制订，唤起了民众对数据安全的强烈关注。

表 6　2019 年数据安全相关立法进程

时间	主要内容
5 月 24 日	网信办发布《网络安全审查办法（征求意见稿）》
5 月 28 日	网信办发布《数据安全管理办法（征求意见稿）》
5 月 31 日	网信办发布《儿童个人信息网络保护规定（征求意见稿）》（已于 8 月 23 日正式公布，自 10 月 1 日开始施行）
6 月 13 日	网信办发布《个人信息出境安全评估办法（征求意见稿）》

来源：中国信息通信研究院

但不可否认的是，从法律法规体系方面来看，我国的数据安全法律法规仍不够完善，呈现出缺乏综合性统一法律、缺乏法律细节解释、保护与发展协调不够等问题。2018 年，十三届全国人大常委会立法规划中的“条件比较成熟、任期内拟提请审议的法律草案”包括了《个人信息保护法》《数据安全法》两部。个人信息和数据保护的综合立法时代即将来临。

（二）数据安全技术助力大数据合规要求落地

数据安全的概念来源于传统信息安全的概念。在传统信息安全中数据是内涵，信息系统是载体，数据安全是整个信息安全的关注重点，信息安全的主要内容是通过安全技术保障数据的秘密性、完整性和可用性。从数据生命周期的角度区分，数据安全技术包括作用于数据采集阶段的敏感数据鉴别发现、数据分类分级标签、数据质量监控；作用于数据存储阶段的数据加密、数据备份容灾；作用于数据处理阶段

的数据脱敏[21]、安全多方计算[22]、联邦学习[23]；作用于数据删除阶段的数据全副本销毁；作用于整个数据生命周期的用户角色权限管理、数据传输校验与加密、数据活动监控审计等等。

当前我国数据安全法律法规重点关注个人信息的保护，大数据行业整体合规也必然将以此作为核心。而在目前的数据安全技术中有为数不少的技术手段瞄准了敏感数据在处理使用中的防护，例如数据脱敏、安全多方计算、联邦学习等等。

在《数据安全管理办法（征求意见稿）》中明确要求，对于个人信息的提供和保存要经过匿名化处理，而数据脱敏技术是实现数据匿名化处理的有效途径。应用静态脱敏技术可以保证数据对外发布不涉及敏感信息，同时在开发、测试环境中保证敏感数据集本身特性不变的情况下能够正常进行挖掘分析；应用动态脱敏技术可以保证在数据服务接口能够实时返回数据请求的同时杜绝敏感数据泄露风险。

安全多方计算和联邦学习等技术能够确保在协同计算中任何一方实际数据不被其他方获得的情况下完成计算任务并获得正确计算结果。应用这些技术能够在有效保护敏感数据以及个人隐私数据不存在泄露风险的同时完成原本需要执行的数据分析、数据挖掘、机器学习等任务。

上述技术是当前最为主流的数据安全保护技术，也是最有利于大数据安全合规落地的数据安全保护技术。其中的各项技术分别具有各

[21] 指对敏感数据通过脱敏规则进行变形从而实现对于敏感数据保护的过程。
[22] 指多个参与实体各自持有秘密输入，各方希望共同完成对某函数的计算，而要求每个参与实体除计算结果外均不能得到其他参与实体的任何输入信息的技术。
[23] 指多个机构在满足用户隐私保护的要求下，进行数据使用和机器学习建模的技术。

自的技术实现方式、应用场景、技术优势和当前存在的问题，具体的对比如下表:

表 7 主要隐私数据保护技术对比

技术类型	数据脱敏	安全多方计算	联邦学习
实现	加密、掩码、k-匿名、l-多样性等	同态加密、秘密分享、不经意传输、混淆电路等	同态加密、混淆电路、可信计算环境等
分类	动态脱敏、静态脱敏	安全两方计算、安全 N 方计算等	横向联邦学习、纵向联邦学习、迁移联邦学习
应用场景	数据对外服务、数据开发、数据挖掘等	数值计算、集合运算、SQL 查询、机器学习等	机器学习
技术优势	应用场景广泛、计算效率高、实现方法多样、可维持部分数据可用性	隐私保护程度高、数据可用性无损失、应用场景较广泛	隐私保护程度高、数据可用性无损失
存在问题	敏感数据辨别、数据脱敏程度与可用度需要平衡	计算性能损失、需要面向场景定制化实现	计算性能损失、可用场景有限

来源：中国信息通信研究院

上述技术均存在多种技术实现方式，不同实现方式可能达到对于隐私数据的不同程度保护，不同的应用场景对于隐私数据的保护程度和可用性也有不同的需求。作为助力实现大数据安全合规落地的主要技术，在实际应用中使用者应根据具体的应用场景选择合适的隐私保护技术以及合适的实现方式，而繁多的实现方式和产品化的功能点区别导致技术使用者具体进行选择时会遇到很大的困难。通过标准对相应隐私保护技术进行规范化，可以有效地应对这种情况[24]。

未来伴随着大数据产业的不断发展，个人信息和数据安全相关法律法规将不断出台，在企业合规方面，应用标准化的数据安全技术是十分有效的合规落地手段。随着公众数据安全意识的提升和技术本身

[24] 可参考《数据流通关键技术白皮书》，中国信息通信研究院，2017 年。

的不断进步完善，数据安全技术将逐渐呈现出规范化、标准化的趋势，参照相关法律法规要求进行相关产品技术标准制定，应用符合相应技术标准的数据安全技术产品，保证对于敏感数据和个人隐私数据的使用合法合规，将成为未来大数据产业合规落地的一大趋势。

（三）数据安全标准规范体系不断完善

相对于法律法规和针对于数据安全技术的标准，在大数据安全保护中，标准和规范也发挥着不可替代的作用。

《信息安全技术 个人信息安全规范》是个人信息保护领域重要的推荐性标准。标准结合国际通用的个人信息和隐私保护理念，提出了“权责一致、目的明确、选择同意、最少够用、公开透明、确保安全、主体参与”七大原则，为企业完善内部个人信息保护制度及实践操作规则提供了更为细致的指引。2019 年 6 月 25 日，该标准修订后的征求意见稿正式发布。

一系列聚焦数据安全的国家标准近年来陆续发布。包括《大数据服务安全能力要求》（GB/T 35274-2017）《大数据安全管理指南》（GB/T 37973-2019）《数据安全能力成熟度模型》（GB/T 37988-2019）《数据交易服务安全要求》（GB/T 37932-2019）等，这些标准对于我国数据安全领域起到了重要的指导作用。

中国通信标准化协会大数据技术标准推进委员会（TC601）推出的《可信数据服务》系列规范将个人信息保护推广到企业数据综合合规。标准针对数据供方和数据流通平台的不同角色身份，从管理流程和管理内容等方面对企业数据合规提出了推荐性建议。规范列举了数

据流通平台提供数据流通服务时，在平台管理、流通参与主体管理、流通品管理、流通过程管理等方面的管理要求和建议，以及数据供方提供数据产品时，在数据产品管理、数据产品供应管理等方面需满足和体现服务能力与服务质量的要求。系列规范已于2019年6月发布。

CAICT 中国信通院

六、大数据发展展望

党的十九届四中全会提出将数据与资本、土地、知识、技术和管理并列作为可参与分配的生产要素，这体现出数据在国民经济运行中变得越来越重要，数据对经济发展、社会生活和国家治理正在产生着根本性、全局性、革命性的影响。

技术方面，我们仍然处在“数据大爆发”的初期，随着 5G、工业互联网的深入发展，将带来更大的“数据洪流”，这就为大数据的存储、分析、管理带来更大的挑战，牵引大数据技术再上新的台阶。硬件与软件的融合、数据与智能的融合将带动大数据技术向异构多模、超大容量、超低时延等方向拓展。

应用方面，大数据行业应用正在从消费端向生产端延伸，从感知型应用向预测型、决策型应用发展。当前，互联网行业已经全面进入“DT 时代”。未来几年，随着各地政务大数据平台和大型企业数据中台的建成，将促进政务、民生与实体经济领域的大数据应用再上新的台阶。

治理方面，随着国家数据安全法律制度的不断完善，各行业的数据治理也将深入推进。数据的采集、使用、共享等环节的乱象得到遏制，数据的安全管理成为各行各业自觉遵守的底线，数据流通与应用的合规性将大幅提升，健康、可持续的大数据发展环境逐步形成。

然而，我国大数据发展也同样面临着诸多问题。例如，大数据原创性的技术和产品尚不足；数据开放共享水平依然较低，跨部门、跨行业的数据流通仍不顺畅，有价值的公共信息资源和商业数据没有充

分流动起来；数据安全管理仍然薄弱，个人信息保护面临新威胁与新风险。这就需要大数据从业者们在大数据理论研究、技术研发、行业应用、安全保护等方面付出更多的努力。

新的时代，新的机遇。我们也看到，大数据与 5G、人工智能、区块链等新一代信息技术的融合发展日益紧密。特别是区块链技术，一方面区块链可以在一定程度上解决数据确权难、数据孤岛严重、数据垄断等“先天病”，另一方面隐私计算技术等大数据技术也反过来促进了区块链技术的完善。在新一代信息技术的共同作用下，我国的数字经济正向着更加互信、共享、均衡的方向发展，数据的“生产关系”正在进一步重塑。

2020 年即将到来，“十三五”规划将圆满收官，“十四五”的号角即将吹响。我们期待，站在新的历史起点上，以大数据为代表的新一代信息技术将对我国制造强国、网络强国和数字中国建设作出更大的贡献。

中国信息通信研究院

地址：北京市海淀区花园北路 52 号

邮政编码：100191

联系电话：13683007576

传真：010-62304980

网址：www.caict.ac.cn