内部资料 仅供参阅

大数据信息快报

2019 年第 1 期（总第 1 期）

北京大学信息管理系
2017 级夜大第九小组编

2019 年 4 月 6 日

本期要目

编辑：
王为臻 印海峰 李怀彬 李勇壮 杨振铎 张兴泽 张朕锋 周在峰 谢鑫酉

目　录

深度关注

关注：2019 年大数据的 10 大发展趋势

【导读】如今，人们寻求获得更多的数据有着充分的理由，因为数据分析推动了数字创新。然而，将这些庞大的数据集转化为可操作的洞察力仍然是一个难题。而那些获得应对强大数据挑战的解决方案的组织将能够更好地从数字创新的成果中获得经济利益。考虑到这个基本前提，以下是2019年应该关注的大数据的10个发展趋势。

1. 数据管理仍然很难

大数据分析有着相当明确的重要思想：找到隐藏在大量数据中的信息模式，训练机器学习模型以发现这些模式，并将这些模型实施到生产中以自动对其进行操作。需要清理数据，并在必要时进行重复。

然而，将这些数据投入生产的现实要比看上去困难得多。对于初学者来说，收集来自不同孤岛的数据很困难，需要提取、转换和加载(ETL)和数据库技能。清理和标记机器学习培训的数据也需要花费大量的时间和费用，特别是在使用深度学习技术时。此外，以安全可靠的方式将这样的系统大规模投入生产需要另外一套技能。

出于这些原因，数据管理仍然是一个巨大的挑战，数据工程师将继续成为大数据团队中最受欢迎的角色之一。

2. 数据孤岛继续激增

这个预测并不困难。在五年前的Hadoop开发热潮中，人们认为可以将所有数据(包括分析和事务工作负载)整合到一个平台上。

出于各种原因，这个想法从未真正实现过。其面临的最大挑战是不同的数据类型具有不同的存储要求。关系数据库、图形数据库、时间序列数据库、HDF和对象存储都有各自的优缺点。如果开发人员将所有数据塞进一个适合所有数据的数据湖中，他们就无法最大限度地发挥其优势。

在某些情况下，将大量数据集中到一个地方确实有意义。例如，像S3这样的云数据存储库为企业提供了灵活且经济高效的存储，而Hadoop仍然是非结构化数据存储和分析的经济高效的存储。但对于大多数公司而言，这些只是必须管理的额外孤岛。当然，它们是重要的孤岛，但它们不是唯一的。

而在缺乏强大集权的情况下，数据仓库将会继续激增。

3. 流媒体分析的突破性的一年

组织处理新数据越快，业务发展就会越好。这是实时分析或流式分析背后的推动力。但组织一直面临的挑战是要真正做到这一点非常困难，而且成本也很高，但随着组织的分析团队的成熟和技术的进步，这种情况正在发生变化。

NewSQL数据库、内存数据网格和专用流分析平台围绕通用功能进行融合，这需要对输入数据进行超快处理，通常使用机器学习模型来自动化决策。

将它与Kafka、Spark和Flink等开源流式框架中的SQL功能相结合，组织就可以在2019年获得真正的进步。

4. 数据治理不善将带来风险

有些人将数据称之为“新石油”，也被称为“新货币”。无论是什么样的比喻，大家都认为数据具有价值，并且如果对此不重视将会带来更大的风险。

欧盟通过去年颁布的GDPR法规阐明了数据治理不善的财务后果。虽然美国还没有类似的法律，但美国公司仍然必须遵守由美国联邦、各州等创建的80个不同的数据制授权法规。

数据泄露正在引发问题。根据Harris Poll公司进行的一项在线调查，2018年有近6000万美国人受到身份盗窃的影响。这比2017年增长了300%，当时只有1500万人表示受到了影响。

大多数组织已经意识到无序发展的大数据时代即将结束。而很多国家

和地区的政府对数据滥用或隐私泄露行为不再容忍。

5. 随着技术的发展，技能也在转变

人力资源通常是大数据项目中的最大成本，因为工作人员最终构建并运行大数据项目，并使其发挥作用。无论使用何种技术，找到具有合适技能的人员对于将数据转化为洞察力至关重要。

而随着技术的进步，技能组合也是如此。在2019年，人们可以看到企业对于神经网络专业人才的巨大需求。在数据科学家(而不是人工智能专家)的技能中，Python仍然在语言中占主导地位，尽管对于R、SAS、Matlab、Scala、Java和C等语言还有很多工作要做。

随着数据治理计划的启动，对数据管理人员的需求将会增加。能够使用核心工具(数据库、Spark、Airflow等)的数据工程师将继续看到他们的机会增长。人们还可以看到企业对机器学习工程师的需求加速增长。

然而，由于自动化数据科学平台的进步和发展，组织的一些工作可以通过数据分析师或“公民数据科学家”来完成，因为众所周知，数据和业务的知识和技能可能会让组织在大数据道路上走得更远，而不是统计和编程。

6. 深度学习变得更加深入

深度学习的发展为人工智能的应用提供了更多的动力，在2019年没有任何减缓的迹象。组织将继续尝试深度学习框架，如TensorFlow、Caffe、Keras、PyTorch和MXnet，以期将大量数据集实现货币化。

组织将扩展深度学习，超越其最初的用例，如计算机视觉和自然语言处理(NLP)，并找到实现强大技术的新的和创造性的方法。大型金融机构已经发现神经网络算法比“传统”机器学习方法更能发现欺诈行为，并且将在2019年继续探索新的用例。

这也将支持对GPU的需求，GPU是培训深度学习模型的首选处理器。目前还不清楚是否有新的处理器类型，包括ASIC、TPU和FPGA。但是，显然还需要更快的培训和推理。

然而，深度学习生态系统将保持相对年轻，缺乏通用平台将使其成为

真正专家的领域。

7. “Special K”扩大了足迹

软件需要运行一些东西。用于提供通用基础的操作系统，但现在开发人员的目标要低一点：Kubernetes。

Kubernetes由Google公司开发，用于管理和协调云中的虚拟化Linux容器，在IT行业中，它已成为大数据生态系统中最热门的技术之一。随着多云和混合部署变得越来越普遍，Kubernetes就是将它们整合在一起的粘合剂。

以前编写Hadoop上运行的软件的大数据软件供应商现在正在编写Kubernetes上运行的软件，这至少让他们进入了前台。支持Kubernetes软件已经成为软件供应商(包括Hadoop供应商)的首要需求。

8. 难以忽视的云计算

云计算的规模越来越大。2018年，全球三大公共云供应商的业务增长率接近50%。云计算供应商提供了一系列大数据工具和技术，更不用说用于存储所有数据的廉价存储，因此用户很难抵御云计算的诱惑。

2019年，小型企业和初创企业将被主要的公共云提供商提供的服务所吸引，这些云计算提供商正在投入巨资建设随时可运行的大数据平台，提供自动化机器学习、分析数据库和实时流分析服务。

即使成本方面并不那么吸引人，大型企业也难以抗拒云计算所带来的好处。然而，将业务锁定在单一云计算供应商，这让大型企业担心面临将所有鸡蛋放在一个篮子中的风险。

9. 新技术将会出现

当今推动创新的许多主要大数据框架和数据库都是由全球网络巨头创建的，并作为开源应用发布。好消息是可能将加快技术创新。

在2019年，大数据从业者在他们的创作中会尽可能保持灵活性。虽然出于性能原因，将应用程序绑定到某项技术可能会很有诱惑力，但是当更好、更快地出现这种情况时，这可能会让组织感到困扰。

尽可能多地保持应用程序“松散耦合但紧密集成”，因为最终必须将其

拆分并重新构建。

10. 智能设备无处不在

如今，智能设备无处不在，并且不断收集数据。而在消费者需求的推动下，智能设备正以惊人的速度增长。智能设备生态系统正在亚马逊Alexa和谷歌智能助理两大领先平台上崭露头角，为消费者提供了将远程访问和人工智能融入从照明、暖通空调系统、门锁、家用电器等各个行业领域的机会。

由于超高速5G无线网络即将推出，消费者将能够与众多设备进行交互，并且无论在哪里，都会提供新的个性化服务。

2019年，大数据将在多个方面取得进展。虽然大数据和人工智能的发展仍然存在大量的技术、法律和道德障碍，但潜在的好处巨大，不容忽视。

（来源：Enterpriseai.news，作者：Alex Woodie，编辑：周在峰）

战略政策

四川省出台政策大力发展大数据产业发展

推进大数据产业聚集发展，是四川省推进大数据发展的重要举措。四川周边省市都在大力发展大数据产业，有些还走得很快。要想让四川省的大数据更具竞争力，就要聚指成拳。所以，近日，四川省印发《大数据产业培育方案》大力支持大数据产业发展。根据方案，四川省将在聚集发展的前提下，对各个城市进行错位布局，协同发展。到2020年，四川省力争建设成为国内一流、中西部地区领先的大数据产业基地。

按照规划，四川将主要在成德绵雅眉泸等市州，聚集发展大数据产业。集聚发展，不仅包括大数据产业的硬件基础，也包括数据中心的建设。按照《培育方案》，下一步四川省还将探索整合利用各个公司的数据中心资

源，打破相关壁垒，提升服务效能。根据《培育方案》，四川省还将推进大数据管理体系研究，出台政务大数据和行业大数据管理办法。

（来源: 四川日报　作者: 唐泽文　编辑：王为臻）

《数字山东发展规划（2018—2022 年）》打造北方大数据高地

为深入贯彻落实习近平总书记视察山东重要讲话、重要指示批示精神，加快实施国家大数据战略，推动数字中国建设，山东省政府印发了《数字山东发展规划（2018—2022年）》，并就数字山东建设作出全面部署安排。

其中在夯实数字山东基础新支撑中提出构建布局合理、规模适度、保障有力、绿色集约的数据中心体系，打造北方地区重要的大数据高地；在做强核心引领产业时要做到突破大数据采集、清洗、存储、挖掘、分析、可视化算法等关键技术。面向重点行业应用需求，开展大数据产业项目试点示范，研发推广一批大数据解决方案及服务。在“互联网+医疗健康”突破行动目标中提出：到2022年，建成覆盖省市县三级的全民健康信息平台，高标准建成国家健康医疗大数据北方中心，远程医疗、分级诊疗、双向转诊信息服务体系基本覆盖各级医疗卫生机构，形成惠及民生的“互联网+医疗健康”服务体系。

（来源：山东省人民政府　编辑：王为臻）

《中国科学院科学数据管理与开放共享办法（试行）》发布

《中国科学院科学数据管理与开放共享办法（试行）》于2019年2月11日正式发布，是中科院实施国家大数据战略的重要举措。

《办法》分为8章，共32条，对中科院科学数据管理与开放共享的总体原则、职责分工、管理要求、保障机制及安全保密等方面的内容进行了明确。重点包括：提出科研项目数据汇交要求，将国家财政性经费支持的各类

科研项目的科学数据纳入项目管理流程，作为项目立项和考核的必要指标；加强科研论文关联数据汇交管理，界定了科研人员、院属期刊和院属法人单位等参与主体职责，建立汇交及审核管理机制，提出了指导性意见；明确科学数据开放共享的原则和主体责任，对院属法人单位、院重大科技基础设施、野外台站等参与主体分别提出了科学数据开放共享的规范性要求；规划院科学数据中心体系，包括总中心、学科中心、所级中心三类，强化科学数据管理与开放共享服务，保障全院科学数据的规范管理与开放共享的可持续发展。

（来源：中国科学院办公厅　编辑：谢鑫酉）

技术动态

Python 开发者年度调研：一半用 JS，2/3 选择 Linux 系统

每年，Python官方都会针对开发者社区做一次年度报告，统计当年的发展情况，并发布调研报告。

2019年，有超过150多个国家的2万多名开发人员加入了这场深入调查，刚刚发布的报告通过7个角度对Python的使用现状、趋势与未来进行了解读。

以下是几个最新的趋势：

1、使用Python作为主语言的开发人员中，有一半的用户也使用JavaScript。Python也经常混搭HTML/CSS，Bash/Shell，SQL，C/C++和Java一起使用。

2、Python用于数据分析比用于Web开发更广泛，数据分析占比58%。

3、84%的用户已经使用Python3，Python2的比例仅为16%。2017年Python3的使用率只有75%。

4、Flask和Django是Web开发人员中流行度最高的框架，两者份额相差无几，但都甩其他Python Web框架“好几条街”。

5、NumPy，Pandas，Matplotlib和SciPy是最受欢迎的数据科学框架和库。

机器学习专用的库如SciKit-Learn，TensorFlow，Keras等也很受欢迎。

6、AWS是Python开发人员最受欢迎的云平台，其次才是Google Cloud Platform，Heroku，DigitalOcean 和Microsoft Azure。

7、在2018年，运维开发者数量明显增加(与2017年相比增加了8个百分点)。在使用Python作为辅助语言的Python用户中，运维已经取代了Web开发成为第一名。

8、PyCharm的专业版和社区版是最受欢迎的Python 开发工具。VS Code已从2017年的7%增加到2018年的16%，成为Python 开发的第二大最受欢迎的编辑器。

9、几乎2/3的Python开发人员选择Linux作为他们开发时的操作系统。

（来源：jetbrains　编辑：李怀彬）

新一代 Angel 正式开源 性能超越 XGBoost 和 Spark

为了迎接对外开源，团队成员对Angel进行了多次重构和升级，可谓是淬火重炼。在此期间，Angel的架构反复改进，性能持续提升。开源前夕，它的性能已经超越了XGBoost和Spark。新一代的Angel，性能更快，功能更强，开发更方便。其改进主要集中在三方面：

生态性：引入PSAgent，支持PS-Service，便于接入其它机器学习框架。

函数性：融合函数式编程特性，自定义psFunc，利于开发复杂算法。

灵活性：支持Spark-on-Angel，Spark无需修改内核，运行于PS模式之上。

新版本的Angel，添加了诸多新功能，最终的目的，就是让算法工程师能更加从容地进行算法优化，融入更多的算法的Trick，让算法的性能，得到了一个飞跃的提升。一个好的开源项目，不但需要有强大的功能和性能，也需要有良好的适配性，能形成好的生态。

超大样本和超高维度的机器学习，在多个真实生产环境中，有着非常普遍的应用场景，这是Angel的切入点，但不是终点和约束，在未来，Angel

还将深入到图计算和深度学习领域，借助开源的力量，做出更多的探索，无论是Wider还是Deeper的模型，Angel都希望能像天使一样，在多个机器学习框架上为它们提速，帮助各个业务提升效果，为AI的发展插上翅膀。

（来源：腾讯大数据　编辑：张朕锋）

腾讯大数据之新一代资源管理与调度平台

云计算、大数据经常意味着需要调动数据中心大量的资源，如何能够快速的匹配合适资源，需要一个聪明的“大脑”。数据平台部的TDW，是腾讯自主研发，支持百PB级的数据存储和计算，提供海量、高效、稳定的大数据平台支撑和决策支持，成为腾讯大数据处理的核心平台。更大规模的集群，更多新的分布式编程框架，更多不同的业务场景，都给这个大脑提出了挑战。

同时，我们也在思考一个并非只为TDW服务的通用资源管理系统。这些价值正是Google Borg十余年来作为secret weapon提供的强大能力，也是Mesos、Corona、Yarn都想追随Borg脚步的原因。

（来源：腾讯大数据　编辑：张朕锋）

行业应用

大数据分析应用于税收征管

据人民邮电报3月29日讯，随着信息技术的飞速发展，税收征管也已经进入大数据时代，由于数据信息化的广泛运用，税务工作既迎来了空前的发展机遇，又遇到了多种挑战，如果不能有效应对，将会给税收工作带来不利影响。挖掘和利用大数据在信息技术飞速发展的时代对税收管理显得尤为重要。笔者建议，为适应税收机关合并后征管需要，在税收分析上引入大

数据，通过多维度分析对比，不断强化税收分析力度，为做实税收征管提供数据支撑。

企业经营“无界性”与税务管理“有界性”的矛盾，迫切需要大数据。互联网推动了企业跨界融合和跨域经营，税务机关无法准确地确认纳税人的经营行为发生地，也就无法准确行使税收管辖权。企业经营的“无界性”，使得应税活动拓展到另一个乃至数个税收管辖权的管辖区域，从而导致税源管理措施失灵，对税务机关传统的层级管理、属地管理方式提出了挑战。

加强信息安全，健全完善制度。网络中存在着一定的无序性和风险隐患，涉税信息的安全问题逐渐突出，不仅给征纳双方带来了不利影响，而且很容易对社会稳定造成影响。针对存在的诸多安全问题，税务机关要积极规范涉税信息采集、存储、应用，加强对税务工作人员的涉税信息安全培训，制定信息管理制度，落实责任制，同时要加大资金投入力度，运用多种技术手段，保障涉税信息的安全。

（来源：中国大数据产业观察　　编辑：李勇壮）

国内将首现金融大数据平台

2019年2月28日，在北京市金融科技促进民营小微企业融资工作会上，北京金控集团宣布发起设立全国首家普惠型金融大数据公司，旨在解决民营和小微企业的融资难题，打造以服务民营和小微企业为主要目标的金融综合服务平台。

北京金控集团为首批五家金控集团监管试点之一，成立于2018年10月19日，并于2019年1月24日核准。由北京国有资本经营管理中心100%持股，注册资本120亿元。银保监会原国际部主任范文仲担任公司的法定代表人及董事长。

关于即将设立的北京金融大数据公司，北京金控集团相关负责人表示，该公司主要整合公共信用信息和社会商业信息，综合运用大数据、云计算

等现代金融科技手段，通过对数据和信息资源的加工、处理，为平台金融机构提供信用评估、风险预警等数据风控服务，解决信息不对称问题。

（来源：北京商报　编辑：谢鑫酉）

过半被调查者有过被大数据“杀熟”经历

3月27日，北京市消协在京发布了《大数据“杀熟”问题调查报告》(以下简称《报告》)。《报告》显示，过半被调查者表示有过被大数据“杀熟”的经历，购物类、在线旅游类和打车类APP或网站大数据“杀熟”最为常见。体验调查发现，部分平台确实存在新老用户同时消费但价格不同的情况，主要是因为新用户拥有优惠券、老用户自动开启了会员资格或推送的商品配置与服务内容存在差异，个别平台涉嫌存在大数据“杀熟”行为。

舆情调查结果显示，88.32%的被调查者认为大数据“杀熟”现象普遍或很普遍，认为大数据“杀熟”现象一般或不普遍的被调查者仅占11.68%，没有被调查者认为大数据“杀熟”现象不存在。此外，有56.92%的被调查者表示自己有过被大数据“杀熟”的经历。网购平台、在线旅游和网约车等被认为消费大数据“杀熟”问题最多。

对此，北京市消协建议，应健全相关法律法规，将大数据“杀熟”行为列入法律治理范围内;创新监管方式方法，建立诚信激励和失信黑名单制度;强化企业诚信自律，同时消费者应提高自我保护意识，在购买商品或服务时尽量做到货比三家，注重个人隐私保护。

（来源：搜狐网　编辑：李勇壮）

对“大数据杀熟”需要创新监管方式

“大数据杀熟”是否存在？其实，不管网络平台如何否认，恐怕都难以改变事实。报告显示，两名体验人员同时通过飞猪旅行网预订麗枫酒店•昌

平体育馆店的同一天高级大床房，老用户的房费不含早餐为291元一间，而新用户的房费不含早餐为286元一间，另享受4元买立减优惠，实际为282元一间。体验结果发现，同一房间新老用户标价不同，优惠也不同，老用户价格高且不享受优惠。这样的情况在其他网络平台也存在。这足以说明“大数据杀熟”的确是存在的，是无法否认的。

消费者虽然清楚地知道自己已经遭遇了“大数据杀熟”，但却无法维权。从调查来看，只有26.72%的被调查者选择向消协或市场监管部门投诉，11.71%的被调查者选择与商家理论、要求赔偿，8.13%的被调查者选择在社交网站或向媒体曝光。剩下的消费者选择忍气吞声或不再在此消费等。这是否说明消费者维权意识不强？答案是否定的。消费者当然也想维权，但维权之路却很艰难。一方面维权举证难，另一方面费时费力且不一定有效果。如此，消费者哪怕拥有最强的维权意识，也很难维权成功。

“大数据杀熟”反映了消费者的知情权、选择权、公平交易权、个人信息受保护的权利未得到充分尊重和有效保障。这需要完善现有法律法规，将数字信息网络中不断涌现的个人信息种类纳入到保护范围内，如网络用户注册信息、搜索记录、定位信息、消费偏好等。最重要的是，创新监管方式方法，采取技术手段和技术设备，建立相应的大数据网上监管平台，对网络平台进行全天候的在线监管，提高对各种隐性大数据利用违法行为的查处能力。同时，要畅道维权渠道，减少维权成本，更要增加处罚，如此，才能倒逼“大数据杀熟”最终退场。

（来源：新文化报　　编辑：李怀彬）

唯品利用用户相似度聚类识别互联网金融欺诈

近些年来，互联网金融发展迅猛，用户逐渐接受并习惯使用消费贷，对现金贷产品的需求也是越来越高。相比于传统金融行业，互联网金融申请、放款、还款流程全部都是线上操作，极大的方便了用户。唯品金融更是实现

了申请秒批。但是由于申请、放款流程的简化，恶意套现、欺诈、逾期不还等时有发生，更有甚者组团套现。用户相似度聚类就是使用dbscan算法将相似用户聚集起来，识别出欺诈团体。

当前大数据时代，技术和业务的发展使得数据的维度与规模实现空前的增长。用户每天在商城和金融产生大量数据，比如：PV、订单、姓名、卡号等。同时为了更好的服务用户，除了用户自己主动填写的数据，客户端还会采集各种数据，比如：WiFi、GPS等。聚类就是基于这些用户信息将相似度高的用户聚集起来。

（来源：唯品金融大数据　编辑：张朕锋）

百度建立数据标注中心角逐 AI 应用市场

2019年3月27日，博鳌亚洲论坛 2019 年年会上，百度副总裁尹世明透露称百度已经在太原建立了数据标注中心。结合其云计算和大数据、人工智能技术等，可以形成一整套产业链，把AI从理想变成现实。

尹世明表示：“必须要给数据做标注，无论是语音的，还是算法的，所以百度现在已经在太原建立了一个巨大的数据标注中心。”

随着AI使用的数据量日益增长，数据标注的需求也逐渐增大，数据标注愈发重要。

何为数据标注？人工智能深度学习是需要根据给定的输入做判断或者预测，其结果的准确性需要大量的数据，因此在数据训练前，必须对大量的数据进行标注。

市面上，数据标注中心主要分为众包和自建工厂两种模式。众包模式主要为头部公司内的数据标注部门，如京东（京东众智）、百度（百度众测）等；自建工厂模式主要为一批数据标注创业公司，如龙猫数据、Testin云测、倍赛BasicFinder、数据堂、阿里数据标注等。

（来源：图灵TOPIA　原创：千鸟　编辑：杨振铎）

国际主流数据库产品性能进入 TB 级分析能力时代

3月25日至28日，在国际大数据会议Strata Data Conference参展的柏睿数据董事长兼CTO刘睿民表示，目前国际主流数据库产品性能已经进入到TB级分析能力时代，核心技术向高性能、高吞吐、高并发、低延时、按需在线灵活扩展等特性进化，要求能够对大规模（即海量）、上百个维度的多源异构数据进行实时加速分析。

和甲骨文、SAP、微软等知名国际数据库企业相比，国内数据库产品在自主技术创新上已经有所突破。和甲骨文、SAP、微软等知名国际数据库企业相比，国内数据库产品在自主技术创新上已经有所突破。其中，在会上亮相的柏睿数据RapidsDB v4.0，在TB级数据毫秒级响应测试中，取得了1秒内在3000亿条数据中匹配唯一1行记录，数据内存空间的占用比例1：1.4的成绩。企业还参与制定了《SQL9075 2018流数据库》和《AI-in-Database库内人工智能》两项数据库国际标准。

数据库是基础软件发展的三大核心之一，目前国产数据库产品目前正进入由基于国际开源产品二次研发向自主研发的关键转型期。数据库核心技术是数字经济发展的关键技术因素之一，加快数据库产品核心技术的自主创新，才能为下一步数字经济与传统经济的融合发展奠定好基础。

（来源：新华网　编辑：杨振铎）

大数据在智慧医疗的应用：把三维光片打造成立体模型

骨架、内脏器官以及外层肌肉脂肪都被立体呈现，用鼠标点击、拖动，三维模型还会朝需要检查的方向转动、放大，任何横截面都有精准的图像……2019年初，一种将X光片变成立体模型的技术在临床中得到了实践。

据医生反馈“这不仅能提高医生手术的精准度，还可以实现模拟手术，

例如一些疑难病症的手术方案，可以利用三维模型讨论手术方案，并通过模型实施，判断哪种手术方案对病人最好。”

通过视频看病，可解决老年人看病频繁却行动不便的难点，患者可在家打开机顶盒，进入医生问诊频道，选择医院、医生，连接后就能通过视频，告知医生自己的症状。除了需要检查的项目，都可以在家完成问诊。

（来源：贵阳日报　作者：王丹丹　编辑：印海峰）

贵阳：大数据助力农业释放新活力

2019年3月下旬，春茶上市，茶场市场的拓展却很艰难，贵州开阳县物联网信息化平台，用“大数据”给出了解决方案：依托贵州智慧农业云平台搭建的物联网采集系统，能通过产品上传的数据，自动生成食品安全追溯体系，给茶叶装上“绿色身份证”，记录品种、出产地、成熟度、环境指纹、品牌以及经过检测、入库、包装、装车运输的各个过程，实现产品的品质追溯。

“扫扫二维码，就能查到茶叶和茶场的各项信息，消费者喝到放心的好茶，能留住不少回头客。”该平台在茶场还建设了农业环境监测设施，通过感应器能实现对茶场环境温度、湿度变化数据监测，并能根据综合分析数据，提供农业指导。

“随着物联网信息化在逐步运用，茶场从田间管理到生产加工再到销售服务都将有新的变革，初步预算，产值至少要增加一倍。”相关负责人说。

贵阳市在逐步试点物联网信息化农业生产的同时，还建成“果蔬生产管理信息服务平台”，实现了“生产数据分析”“产地准出管理”“检测数据自动上传”“标签自动打印”等功能，该平台已覆盖贵阳果、蔬生产企业、基地达109家，涉及10个精品水果品种，涵盖追溯面积10.7万亩。

（来源：人民网　作者：龙章榆　编辑：印海峰）

专家观点

刘权：大数据面临形势与对策

2019年4月4日，工信部赛迪网络安全研究所所长、赛迪区块链研究院院长刘权博士在2019第四届中国网络信息安全峰会上演讲时表示，当前，大数据安全主要表现在几个方面：

网络安全。数据与网络密不可分，针对大数据的网络犯罪行为日益猖獗。目前我国针对大数据的网络安全防护不够，无论是软件还是硬件大多使用国外的产品或技术，容易造成信息泄露；

系统安全。在大数据时代，云平台是大数据汇集和存储的主要载体，云平台数据安全是保证数据安全的重要环节；

终端安全。数据的搜集、存储、访问、传输必不可少地需要借助PC、移动等终端设备，攻击终端设备可能获得操作大数据的权限；

数据安全。大数据时代，看似无用的数据，经过大数据分析技术极有可能转化为由高价值的信息资产。这种信息一旦泄露，将严重威胁个人隐私安全，甚至对国家经济走势、政治稳定产生影响。

大数据平台成为网络攻击的显着目标

大数据时代，作为数据的载体，大型网站、数据中心、云计算中心等大数据平台集聚大量数据，涉及个人隐私、财务等敏感数据，更是业务健康、安全运转的关键，吸引着更多的以商业目的或国家利益为背景的黑客的注意，成为更具吸引力的目标。

数据的大量聚集，使得黑客一次成功的攻击能够获得更多的数据，无形中降低了黑客的进攻成本，增加了“收益率”。

大数据推高信息泄露的风险

海量数据的不断聚集，将促进包括大量的企业运营数据、客户信息、个人的隐私和各种行为的细节记录不断积累，这些集中存储的数据无形中增

加了数据泄露的风险

不法份子可借助大数据平台泄露的数据对其它平台进行“撞库”攻击。由于各企业对数据传输和存储的安全保障能力不一，一旦其中一个较为“脆弱”的数据平台发生数据泄漏，其它“坚固”的平台也将遭受鱼池之殃。

大数据加大网络安全防护的难度

大数据分析技术使攻击者的攻击手段更加丰富。恶意攻击者可以通过收集上网痕迹等信息，获取潜在攻击对象的相关信息，并利用大数据分析技术，对其真实身份、性格、消费习惯、需求等个人信息进行还原，严重威胁个人的隐私和安全。利用数据挖掘、关联分析能够从普通数据中提取大量具有统计意义的信息，可能分析出企业的商业布局，甚至国家的经济走向，进而对企业或者国家发起更具有针对性和精确性的攻击。

传统的防火墙、病毒查杀、入侵检测等安全防护软件不能满足当前需求，防护措施的更新升级速度也无法跟上数据量非线性增长的步伐

大数据也可能成为高级病毒的载体，由于针对大数据等新技术产生的网络威胁的防御体系尚未建立，隐藏在大数据中的病毒和恶意软件难以发现。

大数据对保障基础设施安全和国家主权维护提出新挑战

电信网络甚至工控系统等关键基础设施是大数据发展的基础，大数据安全同样依赖于基础设施的安全，随着经济全球化和供应链全球化的影响，关键基础设施的安全变得日益复杂，一国的基础设施可能同时服务于多个国家，信息经济的高度全球相互依赖性，挑战着原有的国家主权观念。

随着数据价值的不断提高，数据资源成为国家核心战略资产和社会财富，一个国家拥有数据的规模、活性及解释运用的能力将成为综合国力的重要组成部分，对大数据的占有和控制权成为维护国家主权和核心利益的基础。

我国大数据面临的主要问题

法律规范缺失 数据管理缺乏依据

没有对保护本国数据、限制数据跨境流通等做出明确规定，在金融、证券、保险等重要行业在华开展业务的外国企业大量将敏感数据传输、存储至其国外的数据中心，存在不可控风险。

缺乏企业和应用程序关于搜集、存储、分析、应用数据的相关法规，电信、金融、物流等行业个人信息泄露、违规使用情况严重，移动应用多在不必要的情况下采集用户的手机通话记录、短信、地理位置等信息，危及个人财产、生命安全。

产业根基不牢 数据主权面临挑战

大数据安全需要从底层芯片、基础软件到应用分析软件及服务等信息产业全产业链的支撑，我国信息技术起步较晚，大数据相关产业自主能力较差，数据采集、传输、存储、处理等方面技术与国外存在较大差距，在处理芯片、存储设备、大数据软件等方面均存在受制于人的问题。

技术实力较弱 难以应对大数据安全威胁

大数据处理核心技术难以掌控。Hadoop分布式数据处理技术、nosql数据库及流式数据处理技术等分别被国外的Cloudera、IBM以及亚马逊等企业所掌握，国内的数据挖掘、关联分析等大数据关键技术多来自国外，缺乏对大数据技术研发的整体设计框架，与数据安全相关的产品和服务还存在缺口，难以应对大数据应用带来的伴生性安全威胁和传统安全威胁交织的复杂局面。

大数据安全经费投入分散 经费使用效率不高

大数据安全经费投入不足，现有的经费主要用于网络舆情监控等内容审查方面，对大数据安全技术研究、大数据安全产品研发等的支持力度不够，不能满足我国大数据安全发展的需要，影响信息安全产业化进程。

信息安全意识薄弱 大数据安全专业人才供血不足

全民安全意识薄弱，一些部门和人员对信息安全的战略地位认识不到位，“说起来重要，干起来次要”的情况普遍存在。

据Gartner统计，2015年全球新增440万个与大数据相关的工作岗位，并

催生大数据分析师、首席数据官等大数据相关职业。大数据安全发展对需要对数学、统计学、数据分析、机器学习、自然语言处理、网络安全等多方面知识综合掌握，可承担分析和挖掘的复合型人才、高端数据科学家以及管理人才存在很大缺口。

加强大数据安全保障的对策建议

从国家层面上来讲，要尽快制定相关法律法规，明确各方责权；强化制度建设，加强重点领域和行业关键数据的安全监管；加强大数据相关产品、服务管理，建立自主可控信息技术生态体系；加速发展大数据相关技术，建立网络安全纵深防御体系；充分利用区块链技术推动大数据价值实现，确保信息安全；坚持积极防范，构建基于等级保护的大数据纵深防御防护体系架构。

从企业层面来讲，要做到管理和技术并重，全方位提高信息安全防护能力；要做到管理和技术并重，全方位提高信息安全防护能力；要构建大数据等级保护技术框架，确保大数据可信、可控、可管。

（来源：信息化观察网　编辑：周在峰）