云计算与大数据的应用

刘 宁 钟 莲 赵 飞 著

北京工业大学出版社

云计算与大数据的应用

刘 宁 钟 莲 赵 飞 著

北京工业大学出版社

图书在版编目（CIP）数据

云计算与大数据的应用 / 刘宁，钟莲，赵飞著. --
北京 ：北京工业大学出版社，2017.4
ISBN 978-7-5639-5185-7

Ⅰ. ①云… Ⅱ. ①刘… ②钟… ③赵… Ⅲ. ①云计算②数据处理 Ⅳ. ①TP393.027②TP274

中国版本图书馆 CIP 数据核字（2017）第112484号

云计算与大数据的应用

著　　者：刘　宁　钟　莲　赵　飞
责任编辑：张　贤
封面设计：历　程
出版发行：北京工业大学出版社
（北京市朝阳区平乐园 100 号　邮编：100124）
出 版 人：郝　勇
经销单位：全国新华书店
承印单位：北京市迪鑫印刷厂
开　　本：710mm×1000mm　1/16
印　　张：14.5
字　　数：320 千字
版　　次：2018 年 6 月 第 1 版
印　　次：2018 年 6 月 第 1 次印刷
定　　价：52.00 元
标准书号：ISBN 978-7-5639-5185-7

前　言

计算技术的发展经历从合到分，又从分到合的历程，这一发展历程中内在的推动力就是技术。云计算、物联网、社交网络的发展使人类社会的数据发生了变化，社会数据的规模正在以前所未有的速度增长，数据的种类五花八门，对海量异构数据的存储、管理、分析和挖掘成为信息学科的热门领域，大数据技术逐渐进入人们的视野。

计算技术的发展特别是网络技术的发展催生了云计算技术的出现，云计算技术的出现被广泛地认为是信息技术的一次重大变革，大量的与云计算相关的软件和系统架构如雨后春笋般地出现。云计算技术将计算资源、存储资源以及相关各类广义的资源通过网络，以服务的形式提供给资源使用者，改变了传统信息架构中物理资源直接独占使用的模式，甚至从广义上讲只要是通过网络向用户提供服务的信息系统都被称为云计算系统。本书作为云计算与大数据技术的一本综合入门课程，我们一直在思考什么样的人才可以被称为云计算与大数据人才，培养的学生的知识结构是什么样的，云计算与大数据作为一个高速发展的学科，那些知识是必须要了解的。本书并不是对某一项技术的专门介绍，而是希望为学习云计算与大数据技术的同学提供一个完整的知识框架，为今后深入学习奠定基础。

近年来，大数据（big data）一词越来越多地被提及，人们用它来描述和定义信息爆炸时代产生的海量数据，并命名与之相关的技术发展创新。麦肯锡称：“数据，已经渗透到当今每一个行业和业务职能领域，成为重要的生产因素。“大数据”在物理学、生物学、环境生态学等领域以及军事、金融、通信等行业存在已有时日，却因为近年来互联网和信息行业的发展而引起人们关注。数据正在迅速膨胀并变大，它决定着企业的未来发展，虽然很多企业可能还没有意识到数据爆炸性增长带来问题的隐患，但是随着时间的推移，人们将越来越多地意识到数据对企业的重要性。

同样，大数据在医疗与城市发展领域也有着广泛的应用。医疗大数据每天产生的1000 万条数据、一亿余次调阅，是我们亟待数据创新的源动力。同是在信息技术不断更新的过程中，城市的发展始终遵循一定的规律，这种规律隐藏在城市的空间数据和非空间数据中，如何获取、分析并总结出这些规律，对指导城市的发展有着重要的意义。本书编写的目的之一是希望能运用大数据理论和技术分析方法对城市发展的规律进行探索，但到底在城市发展与医疗临床应用中云数据能有多大的作用，如何能更有效的运用，亟待我们去思考、尝试。

最后，由于编者能力有限，书中肯定有不足和遗漏之处，恳请广大同仁朋友批评指正，以便将来做进一步的修订。

编者

理论篇

目录

第 1 章 云计算与大数据概述

本章简要介绍了云计算的历史及其商业和技术驱动力，其次介绍了大数据的时代背景、大数据基本概念、大数据系统以及大数据与企业等方面，最后介绍了云计算与大数据的应用发展。本章很清晰地解释了所有有关的定义，是对于云实践者很有用的参考，图形简单易懂，而且相对独立。大数据的出现帮助商家了解用户、锁定资源、规划生产、做好运营及开展服务。本章讲述了在学习如何以云技术术语进行思考的过程中需要考虑的关键问题，在当今的业务和技术环境中，云计算在连接用户服务、虚拟化资源和应用程序中扮演着中心的角色，本章介绍的内容是十分重要的。

1.1 云计算基础知识

1.1.1 云计算发展简史

1. 简要历史

“云”中计算的发展史可以追溯到效用计算，这个概念是计算机科学家 JohnMcCarthy 在 1961 年公开提出的：

“如果我倡导的计算机能在未来得到使用，那么有一天，计算也可能像电话一样成为公用设施。……计算机应用(computer utility)将成为一种全新的、重要的产业的基础。”

1969 年，ARPANET 项目(Advanced Research Project Agency Network，APRANET，为 Internet 的前身)的首席科学家 Leonard Kleinrock 表示：

“现在，计算机网络还处于初期阶段，但是随着网络的进步和复杂化，我们将可能看到‘计算机应用’的扩展……”

从 20 世纪 90 年代中期开始，普通大众已经开始以各种形式使用基于 Internet 的计算机应用，比如：搜索引擎(Yahoo!、Google)、电子邮件(Hotmail、Gmail)、开放的发布平台(MySpace、Facebook、YouTube)，以及其他类型的社交媒体(Twitter、LinkedIn)。虽然这个服务是以用户为中心的，但是它们普及并且验证了形成现代云计算基础的核心概念。

20 世纪 90 年代后期，Salesforce. com 率先在企业中引入远程提供服务的概念。2002 年，Amazon. com 启用 Amazon Web 服务(Amazon Web Service，AWS)平台，该平台是一套面企业的服务，提供远程配置存储、计算资源以及业务功能。

20 世纪 90 年代早期，在整个网络行业出现了“网络云”或“云”这一术语，但其含义与现在的略有不同。它是指异构公共或半公共网络中数据传输方式派生出的一个抽象层，虽然蜂窝网络也使用“云”这个术语，但是这些网络主要使用分组交换。此时，组网方式支持数据从一个端点(本地网络)传输到“云”(广域网)，然后继续传递到特定端点。由于网络行业仍然引用“云”这个术语，所以，这是相关的，并且被认为是较早采用的奠定效能计算基础的概念。

直到 2006 年，“云计算”这一术语才出现在商业领域。在这个时期，Amazon 推出其弹性计算云(Elastic Compute Cloud，EC2)服务，使得企业通过“租赁”计算容量和处理能力来运行其企业应用程序。同年，Google Apps 也推出了基于浏览器的企业应用服务。三年后，Google 应用引擎(Google App Engine)成为了另一个里程碑。

2. 定义

Gartner 公司在其报告中将云计算放在战略技术领域的前沿，进一步重申了云计算是整个行业的发展趋势。在这份报告中，Gartner 公司将云计算正式定义为：

“……一种计算方式，能通过 Internet 技术将可扩展的和弹性的 IT 能力作为服务交付给外部用户。”

这个定义对 Gartner 公司 2008 年的原始定义做了一点修订，将原来的“大规模可扩展性”修改为“可扩展的和弹性的”。这表明了可扩展性与垂直扩展能力相关的重要性，而不仅仅与规模庞大相关。

Forrester Research 公司将云计算定义为：

“……一种标准化的 IT 性能(服务、软件或者基础设施)，以按使用付费和自助服务方式，通过 Internet 技术进行交付。”

该定义被业界广泛接受，它是由美国国家标准与技术研究院(NIST)制定的。早在 2009 年，MST 就公布了其对云计算的原始定义，随后在 2011 年 9 月，根据进一步评审和企业意见，发布了修订版定义：

“云计算是一种模型，可以实现随时随地、便捷地、按需地从可配置计算资源共享池中获取所需的资源(例如，网络、服务器、存储、应用程序及服务)，资源可以快速供给和释放，使管理的工作量和服务提供者的介入降低至最少。这种云模型由五个基本特征、三种服务模型和四种部署模型构成。”

本书给出了云计算更为简洁的定义：

“云计算是分布式计算的一种特殊形式，它引入效用模型来远程供给可扩展和可测量的资源。”

这个简化定义与之前云计算行业中其他组织定义的版本是一致的。

3. 商业驱动力

在深入探究层层云技术之前，首先要理解导致行业领导者进行创造的动机。本节将要介绍若干激励现代云技术的主要商业驱动力。

这些驱动力从两端影响着云的形成和整个云计算市场，注意到这一点是很重要的。它们促使企业为了支持其自动化需求而采用云计算。同时它们也使得其他组织成为云环境和技术的提供者，创造并满足用户需求。

（1）容量规划

容量规划是确定和满足一个组织未来对 IT 资源、产品和服务需求的过程。这里的“容量”（capacity）是指在一段给定时间内，一个 IT 资源能够提供的最大工作量。IT 资源容量与其需求之间的差异会导致系统效率低下(过度配置)或是无法满足用户需求(配置不足)。容易规划的重点就是将这个差异最小化，以便系统获得预期的效率和性能。

容量规划策略分为如下三种类型：

- 领先策略(Lead Strategy)——根据预期增加 IT 资源的容量。
- 滞后策略(Lag Strategy)——当 IT 资源达到其最大容量时增加资源容量。
- 匹配策略(Match Strategy)——当需求增加时，小幅增加 IT 资源容量。

由于需要估计“使用负载”的变化，因此，容量规划颇具挑战性。在不过度配置基础设施的同时，要不断平衡峰值使用需求。比如，若按照最大使用负载配置 IT 资源，就会出现不合理的资金投入。反之，有限的投资就会导致配置不足，导致由于使用限度降低而出现交易损失和使用受限。

（2）降低成本

IT 成本与业务性能之间的恰好平衡是很难保持的。IT 环境的扩展总是与对其最大使用需求的评估相对应，这可以让不断增加的投资自动支持新的、扩展的业务。大部分所需资金都注入到基础设施的扩建中，这是因为，给定的自动化解决方案的使用潜力总是受限于底层基础设施的处理能力。

需要考虑的成本分为两种：获得新基础设施的成本和保有其所有权的成本。运营开销在 IT 预算中占了相当大一部分，往往超过了前期投资成本。

常见的与基础设施相关的运营成本有如下几种形式：

- 为保证环境正常运行所需的技术人员。
- 引入额外测试和部署周期的更新和补丁。
- 电源和制冷所需的水电费和资金支出。

◆ 维护和加强基础设施资源保护的安全和访问控制措施。

◆ 为跟踪许可证和支持部署安排所需要的行政和财务人员。

持续的内部技术基础设施所有权带来的是沉重责任，这会对企业预算造成多重影响。因此，IT 部门可能成为一个主要的，有时甚至是绝对的花钱部门，它能潜在地抑制企业的反应能力、盈利能力和总体发展。

（3）组织灵活性

企业需要有适应和进步的能力，以便成功应对由于各种因素而导致的变化。组织灵活性是组织对变化响应程度的衡量。

IT 企业常常需要应对行业变化，通常采取的措施是在原来预期或计划的 IT 资源规模上进行扩展。比如，若预算不足，使得原来的容量规划打了折扣，那么即使预见到使用波动不足的基础设施也可能妨碍组织对此作出响应。

在其他情况下，变化的业务需求和优先级也会要求 IT 资源具备更高的可用性和可靠性。比如，即使有足够的基础设施来应对预期的使用波动，也可能由于应用自身的特点降低托管服务器的性能，造成运行异常。由于在基础设施内缺乏可靠性控制，那么，对用户或用户需求的响应可能会导致业务的持续性受到威胁。

从更广泛的范围来说，采用新的或是扩展业务自动化解决方案，所需要的预付投资以及基础设施所有权成本可能会使企业望而却步。企业会勉强接受差强人意的 IT 基础设施质量，因而降低企业满足现实世界需求的能力。

更糟的是，企业在审查其基础设施预算后，可能决定完全不采用自动化解决方案，原因非常简单，那就是企业无法负担该预算。但是，这种无法应对的结果将使得企业无法紧跟市场需求、对抗竞争压力以及实现其战略目标。

4. 技术创新

成熟技术通常是新技术创新的灵感来源，它是新技术创新衍生和建立的实际基础。本节简要介绍了对云计算产生主要影响的前期技术。

（1）集群化

集群是一组互联的独立 IT 资源，以整体形式工作。由于集群固有的冗余和容错特性，当起可用性和可靠性提高时，系统故障率就会降低。

硬件集群的一个必备条件是，它的组件系统由基本相同的硬件和操作系统构成。这样，当一个故障组件被其他组件替代后，集群仍能达到差不多的性能水平。构成集群的组件设备差过专用的高速通信链路来保持同步。

（2）网格计算

计算网格(或“计算的网格”)为计算资源提供了一个平台，使其能组织成一个或多个逻辑池。这些逻辑池统一协调为一个高性能分布式网格，有时也称为“超级虚拟计算机”。网格计算与集群的区别在于，网格系统更加松耦合，更加分散。因此，网格计算系统可以包含异构的，且处于不同地理位置的计算资源，而集群计算系统一般不具备这种特性。

从 20 世纪 90 年代早期开始，网格计算作为计算科学的一部分，其研究工作一直持续着。网格计算项目取得的技术成就影响了云计算平台和机制的方方面面，尤其是通用特性，比如网络接入、资源池、可扩展性和可恢复性。这些特性均以各自特有的形式呈现在网格计算和云计算中。

比如，网格计算以中间件层为基础，这个中间件层是在计算资源上部署的。这些 IT 资源构成一个网格池，实现一系列负载分配和协调功能。中间层可以包含负载均衡逻辑、故障转移控制和自动配置管理，这些都启发了类似的——有些甚至是更复杂的——云计算技术。因此，有些观点认为云计算是早期网格计算的衍生品。

（3）虚拟化

虚拟化是一个技术平台，用于创建 IT 资源的虚拟实例。虚拟化软件层允许物理 IT 资源提供自身的多个虚拟映像，这样多个用户就可以共享它们的底层处理能力。

虚拟化技术出现之前，软件只能被绑定在静态硬件环境中。而虚拟化则打断了这种软硬件之间的依赖性，因为在虚拟化环境中运行的仿真软件可以模拟对硬件的需求。

在一些云特性和云计算机制中能发现现有的虚拟化技术的影子，这些技术启发了云计算的某些核心特性。随着云计算的演化，出现了现代虚拟化技术，这些技术克服了传统虚拟化平台在性能、可靠性和可扩展性等方面的局限性。

作为当代云技术的基础，现代虚拟化技术提供了各种虚拟化类型和技术层次。

（4）技术创新与使能技术

还有其他几个技术也很重要，它们一直都影响着现代云平台技术。这就是云使能技术（cloud-enabling technology）。

- 宽带网络和 Internet 架构
- 数据中心技术
- （现代）虚拟化技术
- Web 技术
- 多租户技术
- 服务技术

在云计算正式出现之前，每种云使能技术都以某种形式存在着。随着云计算的演进，有些技术更加精进了，而有些技术则被重新定义了。

1.1.2 云计算术语

本小节主要阐述一组基础术语，这些术语代表了云及其最基本部件的基本概念和特点。

1. 什么是云

云(cloud)是指一个独特的 IT 环境，其设计目的是为了远程供给可扩展和可测量的 IT 资源，这个术语原来用于比喻 Internet，意为 Internet 在本质上是由网络构成的网络，用于对一组分散的 IT 资源进行远程访问。在云计算正式成为 IT 产业的一部分之前，云符号作为 Internet 的代表，出现在各种基于 Web 架构的规范和主流文献中。现在，同样的符号则专门用于表示云环境的边界，如图 1-1 所示。

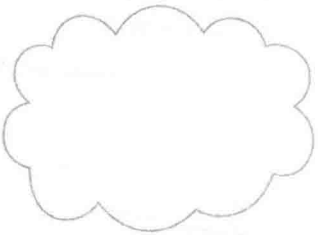

图 1-1 云符号用于表示云环境的边界

区分术语“云”、云符号与 Internet 是非常重要的。作为远程供给 IT 资源的特殊环境，云具有有限的边界。通过 Internet 可以访问到许多单个的云。Internet 提供了对多种 Web 资源的开放接入，与之相比，云通常是私有的，而且对提供的 IT 资源的访问也是需要计量的。

Internet 主要提供了对基于内容的 IT 资源的访问，这些资源是通过万维网发布的。而对于由云环境提供的 IT 资源来说，主要提供的是后端处理能力和对这些能力进行基于用户的访问。另一不关键区别在于，虽然云通常是基于 Internet 协议和技术的，但是它并非必须基于 Web。这里的协议是指一些标准和方法，它们使得计算机能以预先定义好的结构化方式相互通信。而云可以基于任何允许远程访问其 IT 资源的协议。

2. IT资源

IT 资源(IT resource)是指一个与 IT 相关的物理的或虚拟的事物，它既可以是基于软件的，比如虚拟服务器或定制软件程序，也可以是基于硬件的，比如物理服务器或网络设备，如图 1-2 所示。

图 1-2 常见 IT 资源及其对应符号示例

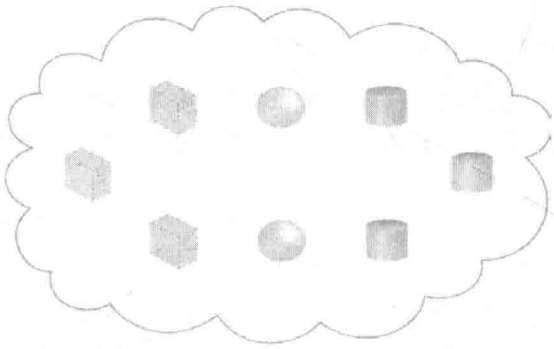

图 1-3 一个包含了 8 个 IT 资源的云，其中有 3 个虚拟服务器、2 个云服务和 3 个存储设备

如图 1-3 所示，用云符号来定义一个云环境的边界，这个云环境容纳并提供了一组 IT 资源。图中所示的这些 IT 资源就被认为是基于云的。

本书含有大量如图 1-3 所示的图示，它们给出了涉及 IT 资源的技术架构和各种交互场景。在学习和使用这些图时，需要注意以下两点：

一个给定云符号边界中画出的 IT 资源并不代表这个云中包含的所有可用 IT 资源。为了说明一个特定的话题，通常只突出显示一部分 IT 资源。

当重点集中在一个问题的某些方面时，就需要特意用抽象图示来表示底层技术架构。这就意味着，在图示中只会显示实际技术的部分细节。

此外，还有些图示中有些 IT 资源在云符号之外，这表示这些资源不是基于云的。

3. 企业内部的

作为一个独特且可以远程访问的环境，云代表了 IT 资源的一种部署方法。处于一个组织边界(并不特指云)中的传统 IT 企业内部承载的 IT 资源被认为是位于 IT 企业内部的，简称为内部的(on-premise)。换句话说，术语“内部的”是指“在一个不基于云的可控的 IT 环境内部的”，它和“基于云的”是对等的，用来对 IT 资源进行限制。一个内部的 IT 资源不可能是基于云的，反之亦然。

有三点需要注意：

- 一个内部的 IT 资源可以访问一个基于云的 IT 资源，并与之交互。
- 一个内部的 IT 资源可以被迁移到云中，从而成为一个基于云的 IT 资源。
- IT 资源既可以冗余部署在内部的环境中，也可以在云环境中。

1.1.3 云计算的特点

为了理解云计算这个概念，只了解一个简单的定义是不够的，我们还需要利用云计算技术的特点来判断一个技术是否是云计算技术。与传统的资源提供方向相比，云计算具有以下特点。

1. 资源池弹性可扩张

云计算系统的一个重要特征就是资源的集中管理和输出，这就是所谓的资源池。从资源低效率的分散使用到资源高效的集约化使用正是云计算的基本特征之一。分散的资源使用方法造成了资源的极大浪费，现在每个人都可能有一到两台自己的计算机，但对这种资源的利用率却非常的低，计算机在大量时间都是在等待状态或是在处理文字数据等低负荷的任务。资源集中起来后资源的利用效率会大大地提高，随着资源需求的不断提高，资源池的弹性化扩张能力成为云计算系统的一个基本要求，云计算系统只有具备了资源的弹性化扩张能力才能有效地应对不断增长的资源需求。大多数云计算系统都能较为方便地实现新资源的加入。

2. 按需提供资源服务

云计算系统带给客户最重要的好处就是敏捷地适应用户对资源不断变化的需求，云计算系统实现按需向用户提供资源能大大节省用户的硬件资源开支，用户不用自己购买并维护大量固定的硬件资源，只需向自己实际消费的资源量来付费。按需提供资源服务使应用开发者在逻辑上可以认为资源池的大小是不受限制的，这就使应用软件的开发者拥有了更大的想象空间和创新空间，更多的有趣应用将在云计算时代被创造出来，应用开发者的主要精力只需要集中在自己的应用上。

3. 虚拟化

现有的云计算平台的重要特点是利用软件来实现硬件资源的虚拟化管理、调度及应用。通过虚拟平台用户使用网络资源、计算资源、数据库资源、硬件资源、存储资源等，与在自己的本地计算机上使用的感觉是一样的，相当于是在操作自己的计算机，而在云计算中利用虚拟化技术可大大降低维护成本和提高资源的利用率。

4. 网络化的资源接入

从最终用户的角度看，基于云计算系统的应用服务通常都是通过网络来提供的，应用开发者将云计算中心的计算、存储等资源封装为不同的应用后往往会通过网络提供给最终的用户。云计算技术必须实现资源的网络化接入才能有效地向应用开发者和最终用户提供资源服务。这就像有了发电厂必须还要有输电线才能将电传送给用户。所以网络技术的发展是推动

云计算技术出现的首要动力。目前一些企业将网络化的软件和硬件都称为云计算，就是因为网络化的资源接入方式是从最终用户角度能看到的云计算的重要特征之一，这些产品的称呼不一定准确但却是对云计算特征的反映。

5. 高可靠性和安全性

用户数据存储在服务器端，而应用程序在服务器端运行，计算由服务器端来处理。所有的服务分布在不同的服务器上，如果什么地方(节点)出问题就在什么地方终止它，另外再启动一个程序或节点，即自动处理失败节点，从而保证了应用和计算的正常进行。

数据被复制到多个服务器节点上有多个副本(备份)，存储在云里的数据即使遇到意外删除或硬件崩溃也不会受到影响。

1.2 大数据技术基础知识

在这个日新月异发展的社会中，人们发现未知领域的规律主要依赖抽样数据、局部数据和片面数据，甚至无法获得真实数据时只能纯粹依赖经验、理论、假设和价值观去认识世界。因此，人们对世界的认识往往是表面的、肤浅的、简单的、扭曲的或者是无知的。然而大数据时代的来临使人类拥有更多的机会和条件在各个领域更深入地获得和使用全面数据、完整数据和系统数据，深入探索现实世界的规律。大数据的出现帮助商家了解用户、锁定资源、规划生产、做好运营及开展服务。

1.2.1　大数据简介

中国庞大的人数和应用市场，其复杂性高并且充满变化，从而成为世界上拥有最复杂的大数据的国家。解决这种由大规模数据引发的问题，探索以大数据为基础的解决方案，是中国产业升级、效率提高的重要手段。因此，解决大数据这一问题不仅提高公司的竞争力，也会提高国家竞争力。

1. 大数据的数据源

近年来，随着信息技术的发展，我国在各个领域产生了海量数据，主要分布如下。

（1）以 BAT 为代表的互联网公司

1）阿里巴巴：目前保存的数据量为近百个拍字节(PB)，90%以上是电商数据、交易数据、用户浏览和点击网页数据、购物数据。

2)百度：2013 年的数据总量接近一千个拍字节(PB)，主要来自中文网、百度推广、百度日志、UGC，由于占有 70%以上的搜索市场份额从而坐拥庞大的搜索数据。

3)腾讯：存储数据经压缩处理后总量在 100PB 左右，数据量月增 10%，主要是大量社交游戏等领域积累的文本、音频、视频和关系类数据。

3)腾讯：存储数据经压缩处理后总量在 100PB 左右，数据量月增 10%，主要是大量社交游戏等领域积累的文本、音频、视频和关系类数据。

（2）电信、金融与保险、电力与石化系统

1)电信：包括用户上网记录、通话、信息、地理位置等。运营商拥有的数据量都在 10PB 以上，年度用户数据增长数十拍字节(PB)。

2)金融与保险：包括开户信息数据、银行网点和在线交易数据、自身运营的数据等。金融系统每年产生数据达数十拍字节(PB)，保险系统数据量也接近拍字节(PB)级别。

3)电力与石化：仅国家电网采集获得的数据总量就达到 10 个拍字节(PB)级别，石化行业、智能水表等每年产生和保存下来的数据量也达到数十拍字节(PB)级别。

（3）公共安全、医疗、交通领域

1)公共安全：在北京，就有 50 万个监控摄像头，每天采集视频数量约 3PB，整个视频监控每年保存下来的数据在数百拍字节(PB)以上。

2)医疗卫生：据了解，整个医疗卫生行业一年能够保存下来的数据就可达到数百拍字节（PB）。

3)交通：航班往返一次就能产生太字节(TB)级别的海量数据；列车、水陆路运输产生的各种视频、文本类数据，每年保存下来的也达到数十拍字节(PB)。

（4）气象与地理、政务与教育等领域

1)气象与地理：中国幅页辽阔，气象局保存的数据为 4～5PB，每年约增数百个太字节(TB)，各种地图和地理位置信息每年约增数十太字节(PB)。

2)政务与教育：北京市政务数据资源网涵盖旅游、教育、交通、医疗等门类，一年上线公布 400 余个数据包。政务数据多为结构化数据。

（5）其他行业

线下商业销售、农林牧渔业、线下餐饮、食品、科研、物流运输等行业数据量还处于积累期，整个体积都不算大，多则达到拍字节(PB)级别，少则几百太字节(TB)，甚至只有数十太字节(TB)级别，但增速很快。

2. 大数据的价值和影响

数量巨大、与微观情境相结合的运行记录信息的最终结果就是大数据。尽管运行记录信息不是大数据的全部，但却应该是以后大数据的主流。目前看得到的金融、电信、航空、电