云计算
——我们真的需要这朵云吗？

现在，云计算（cloud computing）的出现让公司有能力处理海量数据。但是由于云计算目前还存在安全方面的问题，因此只能用它来处理非竞争性或非机密性的数据。

2009年10月，已经被Microsoft公司收购的Danger公司连续4天出现故障，这使得他们的客户——T-Moblie公司的Sidekick手机用户丢失了大量宝贵的数据资料。Daniel Eran Dilger在在线网络杂志——Apple Insider（http://www. appleinsider.com）上撰文指出，该事故进一步放大了“云计算”的弊端。不过，有评论员仍然坚持认为云计算一定会成为处理海量数据的首选运算方式。生物技术行业作为一个充斥着海量数据的行业必定会面临一个抉择，就是该如何面对、利用云计算这个目前被炒爆的新概念。

云计算模型可以让用户借助互联网利用其它地方的计算机（服务器）来运算数据。云计算技术让资金方面捉襟见肘的生物技术行业和缺乏灵活性的大型制药公司在成本、效率和灵活性方面获得了明显的优势。不过，云计算的问题也很明显。虽然尝试过在“中间（媒介）”公司当中设置软件来帮助客户（上述各种专业化的公司）进入云计算系统，并且帮助他们缓冲数据，但是大部分客户都认为这样太不安全。在不久的将来，云计算技术很有希望帮助我们发现早期基因（early stage gene）和不断增多的序列数据的作用。

利用软件充当服务器，aka云计算为大型公司和小型公司提供了方便。

云计算简介

所谓“云”指的就是通过由互联网连接在一起的分布在世界各地的计算机（服务器），而不是像以前那样只能用自己本地的计算机（服务器）来处理数据。因此，有人说云计算是互联网的一种功能延伸。至于云计算这个名称的由来则是因为最开始用“云”的模样来形容互联网的这种组织结构。不过，云计算的最初设想可以追溯到上世纪60年代，那时互联网都还没诞生呢。当时人工智能的开山鼻祖John McCarthy（他同时也是LISP 程序语言的发明人）曾预测，终有一天计算机会被组织起来成为一种公共资源。与其它的公共服务一样，Google、Yahoo、Amazon、Microsoft和AT&T等巨头们也争相推出了云计算服务。

云计算的概念和早几年前出现的网格计算（grid computing）比较相近。所谓网格计算就是将一个大型任务进行分解，然后将这些小型任务分发给众多的服务器进行运算处理。因此，网格计算只适用于对运算要求很高的任务，而且还需要由专家进行操作。一般只有学术机构和公共团体使用网格计算，生物技术行业用得比较少。

相比之下，云计算具有灵活性强和便于操作的优势。不论是仅用一台电脑就可以处理短期任务的新兴小公司，还是资产数十亿的能够处理兆兆（terabytes）级数据的行业巨头都能够使用云计算。

已经使用过云计算的生物技术公司给云计算总结出了几大优势，其中最主要的一条就是能够低成本、高速度地处理大量数据。尤其是小公司可以利用云计算方式将系统管理和数据库管理等工作外包给专业人士运作。

虚拟服务器（virtual servers，即云计算操作过程中利用到的一种服务器）能够帮助那些因缺乏资金而无法购买计算机设备的刚刚成立的小型公司以及对数据处理要求非常高的公司。美国旧金山Penguin 运算公司（该公司主要开发Linux系统）专为生物技术公司构建了一套以云计算技术为基础的大规模平行超级运算集群软件（massively parallel supercomputing cluster），该系统名为Penguin On Demand（POD）。

这种“应需求而生的软件”或者可以称作“利用软件模拟服务器（software as a service）”的技术非常适于普通的运算任务，比如科学运算。Google docs（谷歌办公套件，类似于Microsoft的Office的一套在线办公软件）和Googlemail正在逐步取代我们日常使用的办公系统和电邮系统。生物技术行业也能从这些变革中受益。他们可以解放运算团队，让技术人员专心从事科研工作。

平衡取舍

尽管我们前面介绍了这么多云计算的优势，但是大部分生物技术公司却并不能完全接受这一新生事物。实际上也的确是有一些原因让生物技术公司不会全盘托出他们的数据，比如安全问题、可靠性问题以及知识产权问题等。很多公司都不愿意让他们的敏感数据甚至是任何数据“暴露”在公司防火墙之外。

即使是云计算的大力倡导者，比如英国的咨询师Paul Miller（他同时也是The Cloud of Data的创办者）也非常清楚利用第三方数据中心可能存在的问题。很多云计算服务供应商现在都选择与非常注重数据安全的公司合作，他们希望打造出一套保留有客户公司内部防火墙的“个体云（private clouds）”解决方案。

当政府部门或者药品管理机构加强数据安全，尤其是临床数据安全标准时，上述安全云解决方案就显得尤为重要了。实际上，能够满足这种高标准安全要求的云计算服务供应商才最能吸引到客户。不过万事开头难，有些公司早几年前就已经开始投入公共和个体云服务了，还有一些公司热衷于提供混合解决方案，比如为个体云提供保护或者只处理不那么敏感的数据，但是还有一些公司认为云计算毫无价值。

第一个吃螃蟹的人

早在2003年，就已经有一批生物科技公司开始对云计算技术产生了兴趣。美国新泽西州的制药巨头Merck公司就属于这批勇于第一个吃螃蟹的先导者。当时，Merck公司在基因型、基因表达以及临床试验等方面都远远领先于其它竞争对手。由于需要分析处理兆兆级的数据，Merck公司建立了一套最大型的计算机网路系统，这套系统的规模无论在科研院所里还是在制药公司里都是非常罕见的。构建这套系统的花费达到了数百万美元，最终这套系统达到了拥有10000个节点的规模。但是很快随着第二代测序技术的出现，

这套系统已经无法满足分析处理爆炸式增长的基因数据的要求了。Merck公司也许应该继续扩充计算机网路系统，建立更大规模的网路，选用速度更快的服务器，但是随之而来的各项开支，例如建设新机房、空调设备、水冷设备以及增加员工等等都会压得公司喘不过气来。因此，Merck公司将他们的计算处理需求转移到了美国西雅图的Amazon公司。当时Amazon公司刚刚开始提供云计算服务，因此收费不菲。但是Merck公司很快就发现，这要比自己构建计算机网路划算多了。

到了2009年初，Merck公司关闭了旗下收购的Rosetta制药公司，也随之结束了很多基因组方面的项目，多年来积累的基因组方面的数据被转交到了美国西雅图的一家非盈利性质的公共医学研究组织Sage Bionetworks。Sage获得了Merck/Rosetta Amazon的云计算数据和Merck公司庞大的计算机网路系统，现在他们又在积极地与Microsoft等公司合作，希望在云计算方面取得更大的突破。Schadt现在供职于Sage组织，他同时也是位于美国加利福尼亚州门罗公园的基因组测序公司Pacific Biosciences公司的首席科学官。他对此事评价道：Sage组织最终一定会用到云计算技术，因为世界上已经没有哪一家机构能够仅拥有一套计算机网络系统，就可以应对目前飞速增长的基因组数据了。现在，安全性问题已经不是最需要考虑的问题了，因为针对这些基因组数据的研究还都只是停留在非常初级的阶段。各大公司会自己从Sage组织的数据库中选择他们感兴趣的数据，然后“躲到各自公司防火墙的背后”进行深入的研究。

Pacific Biosciences公司也在从事云计算解决方案方面的工作，不过他们还只是处于初期阶段。Pacific Biosciences公司的第一台商业化的测序仪将于2010年上市。这台测序仪的测序速度可以与著名的Illumina测序仪和罗氏公司的454 FLX测序仪相媲美，不过这还远远不够。Schadt预测未来几年，他们将生产出能够每天获取数千亿序列信息的测序仪。Schadt等人目前的主要工作就是为Pacific Biosciences公司的各位用户提供基因序列数据存储和分析服务。他非常乐观地认为安全问题一定会解决。他认为，Amazon 和Google都已经在数据安全性方面建立起了良好的口碑。各大著名制药企业现在也都敢于将重要数据交给他们进行云计算处理了，这说明云计算服务提供商在某种程度上已经解决了安全性问题。

互相合作

不过正如本文开头提到的Microsoft公司最近出现的问题证明的那样，就算是业界顶尖的公司也不能完全相信。Savvy biotechs公司采用的策略是与好几家云计算服务提供商合作，“中间层（middle-layer）”专业软件就能够帮助他们做到这一点（表1）。

还有一个非常好的例子，就是美国北卡罗来纳州的rPath公司。该公司将各种应用与操作系统、中间设备、图书馆和其它各种所需元件整体打包在一起，这样就可以很方便地在多个云计算服务商和网格计算平台之间进行切换了。他们的这种解决方案能够随时根据需要调整，只需几分钟就能够在Amazon、Rackspace 或者Globus之间进行切换。与rPath公司合作的公司也同样能够进行这种快速切换，而且还能很好地进行数据备份。这样就避免了本文最开始提到的一旦一家服务供应商出了问题，所有数据都丢失了这一悲剧的再度发生。不过有趣的是，只有少数几家生物信息公司开展了这方面的业务。几乎所有的中间层公司都是计算机专业公司，他们的客户都有非常高的数据处理需求度。

有一家生物信息公司紧跟了这股潮流，他们能够提供中间层云计算解决方案。这家公司就是美国西雅图的Geospiza公司。凡是Applied Biosystems公司最新一代测序仪SOLiD测序仪的用户现在都能从这项Geospiza、ABI、Life Technologies和Amazon公司之间的云计算合作项目当中受益。

Geospiza公司的总裁Rob Arnold介绍了这个合作项目的情况：所有研究者都可以将他们的待测序样品交给使用ABI测序仪的测序实验室，待测序工作完成之后序列数据会直接传递给云计算系统进行储存和分析。公司提供的软件系统包括一套电子商业式的前端处理平台，客户（科研人员）们可以直接通过这套系统来选择并且控制他们的数据分析过程。待测样品和测序数据也可以通过一套生命循环处理系统进行从接受订单到得到数据的全程跟踪。所有的数据存储和分析工作都是使用云计算技术完成的，这些“云”既包括Geospiza公司的“个体云”，也包括Amazon公司的“公共云”。

表1 世界著名的几大云计算服务供应商

云计算服务供应商	提供何种服务	主要用户
Amazon公司	提供各种应用及数据储存、分析服务； 提供各种类型的公共云和个体云	Eli Lilly等大型制药公司； ABI等大型生物技术公司； IXICO等小型新兴企业； Sanger中心及EBI等公共科研机构
美国伊利诺伊州的Globus公司	提供网格计算服务	美国能源部等公共机构
Google公司	提供软件虚拟服务器、文件处理系统和电邮系统	Genentech等大型生物技术公司； De novo等小型新兴企业
美国纽约的IBM公司	提供各种应用及数据储存、分析服务	强生制药公司等大型制药公司
Microsoft公司	提供各种应用及数据储存、分析服务	强生制药公司等大型制药公司； Genentech等大型生物技术公司； Spirogen等小型新兴企业； Sage等公共科研机构
英国伦敦的rackspace公司	主要针对小型公司提供各种应用及数据储存、分析服务	Spirogen等小型新兴企业
Star Internet公司	提供数据储存、分析服务和网络服务	Science warehouse等小型新兴企业
美国纽约的Cycle Computing公司	提供网格计算服务	强生制药公司等大型制药公司
Geospiza公司	提供基因测序、数据储存及分析服务	ABI等大型生物技术公司
rPath公司	提供自动化应用服务	美国能源局等公共机构
Solcom公司	提供云计算管理服务和相关软件	Spirogen等小型新兴企业
美国芝加哥的univa uD公司	提供各种云计算管理服务	Pacific Biosciences等小型新兴企业

业界领袖

Amazon公司应该是目前客户首选的云计算服务供应商。这些目前已经用到或者将来可能会用到云计算服务的公司，除了测序公司之外还包括英国剑桥郡的De Novo制药公司，该公司主要利用计算机来辅助药物研发工作；还有英国伦敦的名为IXICO的图像分析公司以及其它一些大型的制药公司，例如Eli Lilly制药公司、Johnson & Johnson制药公司等，还包括一些公共

科研机构，比如基因组学研究所（genomics institutes）等。Amazon公司在云计算服务领域主要有两点优势，即高度的灵活性和目前世界最高标准的安全性。Amazon公司将他们的云计算服务非常形象地命名为“超级灵活的计算云（EC2）”。EC2能够提供全套的各种服务。比如小型公司或科研团队每小时只用花费几美分（甚至还能刷卡）就能租用到Amazon公司的“虚拟Linux盒（virtual Linux box）”云计算业务及相关软件。跨国巨头们也能从Amazon公司得到他们想要的服务。他们可以从Amazon公司获得一套虚拟超级计算机系统，而不需要自己为建立这样一套系统来支付高昂的费用。Eli Lilly公司就使用了Amazon公司的虚拟个人网络系统（Virtual Private Network）。这套系统可以让Eli Lilly公司在本公司防火墙及其它安全措施的保护下享用云计算服务。

Amazon公司的云计算服务所具备的高安全性（这一点尤其能够满足生物技术公司的需要）甚至都符合了《美国健康保险流通与责任法案》（HIPAA）的要求。《美国健康保险流通与责任法案》除了能够规范健康保险政策之外，还赋予了与个体健康相关的数据隐私权和保密权。这套复杂的规章制度还没有被美国的医疗研究人员普遍接受，但是像Amazon公司这样一家并不非常关注于健康事业的公司都能够很好地遵守该法案的规定，说明他们非常看重生物技术行业的数据安全问题。“Amazon公司已经就云计算安全问题发布了白皮书，书中承诺他们将完全遵守《美国健康保险法案》的一切规定。这也是我们为什么选择EC2的原因。”IXICO公司的IT咨询师Norman Taylor这样解释道。

IXICO公司的核心技术（业务）是为临床试验提供图像分析和数据管理服务。公司客户包括世界各地的大型制药企业、生物技术公司和学术科研机构。Taylor介绍说：“我们公司的技术开发人员目前正在测试一款名为Trial Tracker的最新软件，这是我们最新的在线数据管理系统，它是基于Amazon公司的云计算服务而设计开发的。我们相信这套解决方案将能够满足我们部分客户的要求，不过我们也知道，云计算服务并不能满足所有人的需要。”

顺应时代发展

虽然上面介绍的这些例子都表明各种规模的制药公司都需要用到云计算服务，这些公司也可能不需要云计算服务的超级计算能力。英国剑桥一家新成立的抗生素制药公司Prolysis公司就认为他们对数据存储的需求并不是很大。而德国汉堡的一家名为Evotec的药品研发公司也因为云计算技术固有的不安全性而拒绝使用该服务。

不过，就算上面介绍的这些公司不使用云计算服务，他们也不能完全违背时代发展的进程。恐怕只有很小一部分研究人员可能会说他们永远也不会使用公共的生物信息资源，或者可能还有更少的人也许永远都不会用到实验耗材。要知道这两项服务现在都是由使用着云计算服务的公司和科研机构来提供的。现在，由英国Wellcome基金会桑格研究所（Wellcome Trust Sanger Institute）和德国海德堡的欧洲分子生物学实验室欧洲生物信息学研究所（EMBL-EBI）共同合作开展的真核生物基因组数据库项目也使用了Amazon公司的公共云服务。EMBL-EBI核酸数据库的负责人Ewan Birney指出，他们的数据库一直以来对所有人都是免费的，包括公司。EMBL-EBI和桑格研究所的其它一些数据库也将陆续在各种公共云和个体云计算服务中使用。

上面介绍的这些公共基因组学研究机构对数据转移和处理的需求要比各大制药公司大得多，而且这种需求还在不断地增长。比如由桑格研究所的Richard Durbin和美国哈佛大学的David Altshuler共同领导的千人基因组计划（1000 Genomes Project，该计划旨在对人类基因变

异情况进行深入的分类研究）就会将我们对数据处理的能力向上提升一个等级。各大云计算服务供应商的出现为这些科研机构节省了大笔资金和大把时间。桑格研究所表示，他们可以把Illumina测序仪得到的序列数据完全交给云计算服务供应商去处理，相比自己处理既省钱又省时。因为这些数据都属于对公共开放的数据，因此相关的安全问题也要少很多。

不过即便如此，桑格研究所还是面临了一些有关公共云计算服务方面的问题，最大的就是数据传输的问题。现有的互联网还不适合传输兆兆级的数据，有时用硬盘和Fedex快递反而更快更便宜。Simon Twigger是美国威斯康辛医学院的一家机构，他们主要利用Amazon公司的云计算服务通过一套虚拟蛋白质组学数据分析集设备（ViPDAC）来提供蛋白质组学数据分析服务。这套ViPDAC也是对所有人包括商业用户免费开放的，Twigger不会对用户的使用情况进行任何记录。

那些为生物技术公司提供实验耗材的公司也都在积极地参与到云计算领域。英国利兹的Science Warehouse公司就是一家以网络为基础的采购公司，他们主要从事科学相关的业务，客户都是学术科研机构，不过最近生物技术公司类的客户也越来越多。Science Warehouse公司最近为他们公司的数据库开发了一套云计算系统，他们在开发这套系统时就是租用了英国Gloucester的Star Internet公司的服务器，然后在自己公司内部完成了整套系统的开发工作。Science Warehouse公司认为，获得信息是非常敏感的一件事。虽然没有制药企业那么敏感，但是他们对此也相当慎重，投入了大量的资源尽可能地保证我们自己软件的安全。

做正确选择

如果我们能够用谨慎、乐观的态度来看待云计算这一新生事物的话，估计应该大部分公司的商业模式里可能都会包含云计算这项业务。目前看来，似乎没有一个云计算商业模式能够适用于所有的生物技术公司，也不会有太多的企业经理人和科学家能够完全忽视云计算技术。不过，也没有人能像Creative Commons公司的科学副总裁John Wilbanks那样乐观。Wilbanks认为云计算技术正在快速普及，因为他说：“这项技术使得网络变得更加实用”，各种享有共同标准的数据也在不断进入公共领域。

现实情况是，生物技术行业正在引领“混合经济（mixed economy）”的发展，他们将公共云、个体云和传统的本地计算机系统混合起来。虽然各个行业对这个问题的重视程度不同，不过安全问题仍然是一个不容忽视的问题。强生制药公司的Franckowiak在经过了仔细的研究之后总结说，如果做出了明智的选择，那么云计算技术是能够部分解决强生公司的计算需求的。他还说：“云计算可以解决繁重的企业内部计算问题，但是至少目前还不能利用云计算来处理非常珍贵的数据资料，因为这样做的风险太大了。”

原文检索：
Clare Sansom. (2010) Up in a cloud? *Nature Biotechnology*, 28(1):13-15.

YORK/编译

注：本文观点仅代表原作者个人意见，不代表本刊立场！