《代码英雄》第二季（6）：数据大爆炸

代码英雄讲述了开发人员、程序员、黑客、极客和开源反叛者如何彻底改变技术前景的真实史诗。

什么是《代码英雄》

Command Line Heroes

代 码 英 雄是世界领先的企业开源软件解决方案供应商红帽（Red Hat）精心制作的原创音频播客，讲述开发人员、程序员、黑客、极客和开源反叛者如何彻底改变技术前景的真实史诗。该音频博客邀请到了谷歌、NASA 等重量级企业的众多技术大牛共同讲述开源、操作系统、容器、DevOps、混合云等发展过程中的动人故事。

本文是《代码英雄》系列播客第二季（6）：数据大爆炸的音频脚本。

导语：大数据将有助于解决大问题：我们如何种植粮食、如何向需要的人运送物资、如何治疗疾病。但首先，我们需要弄清楚如何处理它。

现代生活充满了相互联系的组件。我们现在一天产生的数据比几千年来的数据还要多。 Kenneth Cukier 解释了数据是如何发生变化的，以及它是如何开始改变我们的。Ellen Grant 博士告诉我们波士顿儿童医院是如何使用开源软件将堆积如山的数据转化为个性化治疗方法。Sage Weil 则分享了 Ceph 的可扩展和弹性云存储如何帮助我们管理数据洪流。

收集信息是了解我们周围世界的关键。大数据正在帮助我们拓展永不停歇的探索使命。

00:00:03 - Saron Yitbarek：

如果你把从人类历史早期到 2003 年创建的所有数据计算在内，你将得到大约 500 万 GB 的数据。我们昨天创建了多少 GB 数据？

00:00:15 - 问卷调查答案 1：

哦，天哪，10 万。

00:00:21 - 问卷调查答案 2：

可能是 500 万 GB 数据。

00:00:23 - 问卷调查答案 3：

我们在昨天一天之内创建了多少 GB 的数据？1000 万 GB 数据？

00:00:31 - 问卷调查答案 4：

我不太知道，可能是 200 万 GB 数据？

00:00:36 - 问卷调查答案 5：

也许一天就有 100 万 GB 数据？

00:00:40 - Saron Yitbarek：

答案是？超过 25 亿 GB 数据！

00:00:44 - 问卷调查答案 1：

哇哦。

00:00:44 - 问卷调查答案 2：

25 亿？

00:00:45 - 问卷调查答案 3：

所以，我们已经打破了世界纪录。

00:00:45 - 问卷调查答案 4：

那可真是很多个 G 啊。

00:00:45 - 问卷调查答案 5：

我都不敢相信那有这么多的数据。

00:00:52 - Saron Yitbarek：

在 2016 年，我们的年度在线数据流量首次超过了 1ZB。一个 ZB 是 1000^7 字节。好，记住这个数字了吗？现在把它乘以 3，因为那是我们将在 2021 年拥有的数据量大小。

00:01:10：

我知道，大脑不是以 ZB 为单位进行思考的，但请你至少暂时记住这个数。 我们的 IP 流量将在五年内翻三番。 这是数据的洪流，而我们正处于其中。

00:01:24：

在刚过去的一分钟里，人们发出了 1600 万条短信；与此同时，在我说出这句话的时间里，谷歌处理了 20 万条搜索。

00:01:37：

如果我们能在数据洪流来临时做好准备、站稳脚跟，那么隐藏在其中的模式、答案和秘密可以极大地改善我们的生活。

00:01:50：

我是 Saron Yitbarek，这里是《代码英雄》，一款红帽公司原创的播客节目。浪潮近在眼前。这里是第二季第六集，数据大爆炸。

00:02:17：

我们如何处理如此大量的数据？ 采集到这些数据后，我们如何利用它们？ 大数据将为我们解决一些最复杂的问题：

00:02:29：

如何管理交通、如何种植粮食、如何向需要的人提供物资，但这一切的前提是，我们必须弄清楚该怎么使用这些数据、以及该怎么在短得不能再短的时间内完成对它们的处理。

00:02:43 - Kenneth Cukier：

通过获取更多的数据，我们可以深入到这些子群体、这些细节，而这是我们以前从来没有过的方式。

00:02:53 - Saron Yitbarek：

Kenneth Cukier 是《The Economist经济学人》的高级编辑，他也和我都在科技播客《Babbage》里。

00:03:01 - Kenneth Cukier：

这并不是说我们以前无法收集数据。我们可以，但这真的、真的很昂贵。真正的革命性突破是，我们使数据搜集变得十分容易。

00:03:10：

现在收集数据的成本极低，而且处理起来也超级简单，因为都是由电脑完成的。这已经成为我们这个时代的巨大革命，它可能是现代生活最显著的特征，在未来几十年甚至下个世纪都会如此。这也是大数据如此重要的原因。

00:03:33 - Saron Yitbarek：

历史可以提醒我们这种变化是翻天覆地的。想想看，4000 年前，我们把所有的数据都刻在了干泥板上。

00:03:46 - Kenneth Cukier：

这些黏土盘很重。黏土盘被烤干后，刻在其中的数据就无法更改。从古至今，信息处理、存储、传输、创建的方式都发生了变化，对吗？

00:04:04 - Saron Yitbarek：

时代巨变。大约在 1450 年，印刷机的发明带来了第一次信息革命。今天，我们也迎来了一场革命。

00:04:16 - Kenneth Cukier：

现在的存储介质很轻巧。信息的修改变得极度简单，借助现有的处理器，我们只需要使用删除键就能修改我们所拥有的信息实例，无论那是储存在磁带上还是晶体管里。我们可以以光速传输数据，不用携带什么黏土盘。

00:04:37 - Saron Yitbarek：

在 15 世纪，借助印刷机，大量的数据得以传播。这些知识提升了人们对事物的认识，并促成了启蒙运动。

00:04:45：

今天，大数据可以再次提升我们的知识水平，但我们必须要想办法充分利用这些数据。唯有修好大坝、备好轮机，才能让浪潮为人所用。

00:05:00 - Kenneth Cukier：

当下，人们远没有做到对数据的充分利用。这一点非常重要，因为我们已经看到，数据中存在这种潜在的价值，而收集、存储和处理数据的成本在近百年来，乃至于近十年来，已经显著地降低了。

00:05:22 - Kenneth Cukier：

这很振奋人心。但问题是，我们在文化上、在我们的组织流程上，甚至我们的 CFO 和 CIO 们拨给相关方面的预算中，并不重视这种价值，

00:05:35 - Saron Yitbarek：

想着这种事肯定让人极度沮丧。启蒙运动在敲门，却无人应答。然而，我们不回答的部分原因是：门后到底是谁？这些数据能带来什么？

00:05:51：

Kenneth 认为，某些公司不采用大数据，乃是因为它太过于新奇。

00:05:56 - Kenneth Cukier：

一旦你收集了大量数据后，你能拿它干什么？我就直说吧，只有傻子才会以为自己知道。你绝对无法设想你今天收集到的数据明天能拿来做什么用。

00:06:12：

最重要的是要有数据，并以开放的思想对待所有可以使用的方式。

00:06:18 - Saron Yitbarek：

如果我们按照 Kenneth 说的那样以正确的方式对待大数据，我们将会发现一切的全新可能性。这将是一个人人 —— 而不只是数据科学家 —— 都能洞察趋势、分析细节的世界。

00:06:33 - Kenneth Cukier：

如果我们能意识到，这个世界是可以通过收集经验证据来理解、改变和改善的，并且可以用一种自动化的方式进行改善，我们将会得到看待它的全新角度。我个人认为，现如今，在世界各地，上至政策制定者、下至星巴克咖啡师，都在经历这种引人深思的文化上或心理上的变化。

00:07:00：

各行各业的人都有点数据基因，就像是被感染了似的。现在，无论他们专注于什么方面，他们都以大数据的方式思考。

00:07:15 - Saron Yitbarek：

Kenneth Cukier 给我们讲了一个简短的故事来展现这种新数据思维方式的力量。微软的一些研究人员开始着手研究胰腺癌问题。

00:07:27：

人们发现胰腺癌往往为时已晚，早期发现可以挽救生命。因此，研究人员开始询问这些患者，在开始搜索有关胰腺癌的信息之前几个月，他们搜索了什么？ 而早在发现前数年，他们又搜索了什么？

00:07:46：

研究人员开始寻找埋藏在所有搜索数据中的线索和模式。

00:07:54 - Kenneth Cukier：

他们有了重大发现。通过分析患者在最终开始搜索“胰腺癌”之前的这段时间中所搜索的关键词，他们识别出了一套规律，可以非常准确地预测搜索者是否患有胰腺癌。

00:08:09：

在这里，我们能学到一点：想象力与数据中潜在规律的结合，是可以挽救生命的。他们现在所要做的就是找到一种方法，通过方法来解释这一发现，这样当人们在搜索这些术语时，他们可以以一种微妙的方式干预，说，“你可能要去诊所检查一下。”

00:08:29：

像这样使用数据，就能救人于水火之中。

00:08:37 - Saron Yitbarek：

研究人员偶然发现的是一种新的癌症筛查方式，通过这种方法，患者可以提前一个月得知自己可能患癌。利用数据不仅仅是一个利润或效率最大化的问题。

00:08:52：

它的意义远不止于此。对于人类而言，这些数据中确确实实存在着大量的潜在利好。抗拒使用大数据可能只是自欺欺人。接下来，我们要关注的是，这场将数据投入工作的持久战。

00:09:18：

哈佛医学院的波士顿儿童医院去年完成了 26000 多台手术，进行约 25 万人次的儿童放射检查。

00:09:31：

医护人员的表现令人称道，但有一个巨大的障碍挡在他们面前。

00:09:37 - Ellen Grant：

在医院的环境中，尤其是作为医生，我们经常会遇到难以获取数据的问题。

00:09:45 - Saron Yitbarek：

这位是 Ellen Grant 医生，她是波士顿儿童医院的儿科神经放射科医生，她在诊疗时依靠访问数据和分析医学图像。

00:09:56 - Ellen Grant：

如果没有专门设置的环境，想要从 packs 里存储的图像进行额外的数据分析绝非易事。当你在一个只提供了普通的医院电脑的读片室里时，要做到这一点并不容易。

00:10:14：

获取数据实际上是有障碍的。

00:10:17 - Saron Yitbarek：

其实许多医院都会大量抛弃数据，因为存储它们的成本实在过于高昂。这部分数据就像这样丢失了。像 Grant 这样的放射科医生可能是第一批因为数据实在太多而感到沮丧的医务人员。

00:10:33：

当医院走向数字化后，他们开始创造大量的数据，很快，这个量就大到无法处理了。

00:10:41 - Ellen Grant：

我，作为一名临床医生，在读片室里的时候希望能将所有复杂的分析工作在研究环境中做完。但我无法随便就从 packs 中拿出来图像，拿到一些可以进行分析的地方，再拿回到我手里。

00:10:59 - Saron Yitbarek：

顺便说一句，packs 就是医院存储其图像的数据仓库。Grant 医生知道有一些工具可以让这些图像 packs 发挥更大的功能，但成本太高。

00:11:12 - Ellen Grant：

随着机器学习和 AI 时代的到来，数据的生产量将会日渐加大，我们会需要更多计算资源来进行这类大规模的数据库分析。

00:11:27 - Saron Yitbarek：

数据已经堆积如山了，但处理能力却没有相称的增长。在这一前提下，对数据的彻底处理将变得遥不可及。而复杂、昂贵的超级计算机并不是医院的选择。

00:11:41：

Grant 医生深感沮丧。

00:11:44 - Ellen Grant：

我们能不能想出一个更好的办法，让我把数据拿到这里来，分析一下，然后放回去，这样我就可以在会诊的时候，一边解释临床图像，一边把分析做完，因为我希望可以在会诊上展示数据，在此同时进行快速分析。

00:11:56：

我可不想在不同的电脑和存储器之间把这些数据挪来挪去，这不是我的工作。我的工作是理解非常复杂的医学疾病，并把相关的事实真相记在脑子里。

00:12:10：

我想专注于我的技术领域，在此同时利用计算机领域的新兴技术；而不必这方面过于深入钻研。

00:12:21 - Saron Yitbarek：

Grant 和世界各地的放射科医生们需要的是一种方法，只要点击图像就能运行详细分析，并让这一切都发生在云端，这样医院就不必建立自己的服务器场地，也不必把医务人员变成程序员。

00:12:40：

他们需要一种方法来使他们的数据尽可能地拯救生命。这正是 Grant 医生和几位代码英雄决定去做的事。

00:12:55：

Grant 在波士顿儿童医院的团队正在与红帽和马萨诸塞州开放云（MOC）合作。关于 MOC 的更多内容稍后再说。首先，我们需要请出 Rudolph Pienaar，他是医院的一名生物化学工程师，来描述一下他们的解决方案。它是一个开源的、基于容器的成像平台。

00:13:15 - Saron Yitbarek：

它完全是在云端运行的，所以你不受医院本身计算能力的限制。他们称这一作品为 ChRIS。

00:13:24 - Rudolph Pienaar：

ChRIS 有一个后台数据库，其实就是一个 Django Python 机器。它可以跟踪用户，并跟踪这些用户使用过的数据以及分析结果。

00:13:35：

围绕这个数据库，有大量的服务群，这些服务都是作为自己的实例存在于容器中。它们处理与医院资源的通信，比如与医院数据库的通信。这些服务从资源中提取复杂的数据，将其推送给云端的、或者另一个实验室的、或者别的什么地方的其他服务处理。在计算数据的地方，有 Kubernetes 之类的编排服务，以及你需要使用的分析程序。数据处理结束之后，结果就会被发送回来。

00:14:11 - Saron Yitbarek：

对于 Grant 医生来说，ChRIS 成像平台是一种让数据活起来的方法。更重要的是，这种数据处理方式能让她成为更好的医生。

00:14:21 - Ellen Grant：

优秀的医生之所以优秀，是因为他们在一生中积累了丰富的从业经验。如果我们能把这一点融入到数据分析中，以此来获得更多的信息，我们就能知道得更多，并更有效地整合这些经验。

00:14:39：

例如，我对特定病患的特定受伤方式的认识，取决于我的从医经验和对这些经验的整体理解。

00:14:52：

现在，我可以根据真实数据创建受伤症状分布的概率图，并将其公之于众；我也可以寻找有相似模式的患者，并告诉他们在接受治疗时，什么对他们最有效，以便更接近精准医疗。

00:15:10：

整合大量的数据，尝试探索我们过去的知识，并尽你所能，点明治疗病人的最佳方式。

00:15:21 - Saron Yitbarek：

这对被送到医院的孩子意味着什么？Grant 医生说，ChRIS 平台能提供更有针对性的诊断和更个性化的护理。

00:15:31 - Ellen Grant：

如果我们拥有更复杂的数据库，我们就能更好地理解信息之间繁杂的相互作用，因此就能更好地指导每个患者。我认为 ChRIS 就像是我进入超级大脑的接口，它能让我比平时更聪明，因为我不能一次把所有数据保存在我的大脑中。

00:15:53 - Saron Yitbarek：

当赌注如此沉重时，我们要突破人类大脑的极限。这位是 Máirín Duffy。她是红帽团队中的设计师，她让 ChRIS 成为现实，而根据个人经验，她知道这件事其中的风险。

00:16:15 - Máirín Duffy：

我父亲中风了，所以我一直作为病人家属等待医疗技术诊断，因为当一个人中风并被送到医院之后，医务人员必须弄清楚是哪种类型的中风。根据中风类型，有不同的治疗方法。

00:16:31：

如果使用了错误的治疗方案，就可能发生极其糟糕的事。所以，在这种情况下，你能越快地把病人送来做核磁共振，就能越快地得到治疗方案。速度越快就有可能挽救他们的生命。

00:16:43：

想想看，仅仅是把图像处理从云端推送出来，并行化处理，就能让它快很多。这样就能将这个过程从几小时、几天，缩短到几分钟。

00:16:55 - Saron Yitbarek：

医学可能正迎来一个新的拐点。一个不是由药理学驱动，而是由计算机科学驱动的拐点。另外，想想像 ChRIS 这种东西的拓展性。

00:17:08：

发展中国家的医生也可以受益于波士顿儿童医院的专业知识和数据集。任何有手机服务的人都可以通过网络访问能够拯救生命的数据和计算结果。

00:17:24：

除了医学，很多其他领域也可能出现类似的拐点。但前提是，人们得知道如何从自己的数据中找到隐藏信息。为了做到这一点，他们需要探索一个全新的计算领域。

00:17:46：

世界各地的人们都在学习如何利用数据。就像在波士顿儿童医院一样，将数据洪流导向目标。

00:17:56：

换句话说，我们在处理这些数据。但我们之所以能做到这一点，是因为新一代的云计算使之成为可能。

00:18:11：

对于像 ChRIS 这样的平台来说，一个关键因素是基于云计算的新型存储方式。请记住，很多医院都会把收集到的数据扔掉，因为他们根本无法容纳所有数据。

00:18:25：

这就是我想重点讨论的数据泛滥的最后一块拼图：存储解决方案。对于 ChRIS 来说，存储解决方案是一个叫 Ceph 的开源项目。它使用的马萨诸塞州开放云，就基于 Ceph。

00:18:45：

我和 Ceph 的创建者 Sage Weil 聊了聊，想了解更多关于像波士顿儿童医院这样的地方是如何在闪电般的时间内处理海量数据的。以下是我与 Sage 的对话。我认为，第一个重要问题是，什么是 Ceph，它能做什么？

00:19:05 - Sage Weil：

当然，Ceph 是一个由软件定义的存储系统，它允许你提供可靠的存储服务，并在不可靠的硬件上提供各种协议。

00:19:14：

它的设计从开始就是满足可扩展性，所以你可以拥有非常非常大的存储系统、非常大的数据集。于此同时，系统对硬件故障和网络故障有

优秀的容忍性，所以即使出现了一些这类问题，存储中的数据仍然不会变得难于访问。

00:19:29 - Saron Yitbarek：

现在，数据太多了。

00:19:31 - Sage Weil：

是的。

00:19:33 - Saron Yitbarek：

如此大的工作量。要处理的东西实在是太多了。你认为这个解决方案出现得是时候吗？

00:19:39 - Sage Weil：

是的，肯定是这样。在当时，行业中这方面的严重不足是显而易见的。没有开源的解决方案可以解决可扩展的存储问题。所以，我们显然得造个轮子。

00:19:53 - Saron Yitbarek：

考虑到我们每天要处理的数据量，以及它将来只会越来越多、越来越难管理的事实，你认为当今该怎么做才能解决这种日益增长的需求？

00:20:09 - Sage Weil：

我认为有几方面。一方面，有令人难以置信的数据量正在产生，所以你需要可扩展的系统。它不仅可以在硬件和数据规模上进行扩展，而且，它的管理成本应该是一定的，至少应该基本固定。

00:20:25 - Saron Yitbarek：

嗯。

00:20:26 - Sage Weil：

你不会想就为每多 10PB 存储空间或类似的东西就多雇一个员工吧？我认为这套系统在运维上也必须可扩展。

00:20:33 - Saron Yitbarek：

是的。

00:20:35 - Sage Weil：

这是其中的一部分。我认为，人们利用存储空间的方式也在改变。一开始，都是文件存储，然后我们有了虚拟机的块存储，我觉得对象存储在某种程度上是行业的重要趋势。

00:20:51：

我认为，下一个阶段的目标并不局限于提供一个对象存储端点，并将数据存储在集群中；我们需要将解决方案进一步升级，好让它能管理集群的集群，抑或是对分布于不同地理位置的云空间及私有数据中心储存空间中的数据进行管理。

00:21:13：

例如说，你现在将数据写入一个位置，随着时间的推移，你可能会想将数据分层到到其他位置，因为它更便宜、或者服务器离你更近；或者，一旦数据太老、不会频繁使用了，你就需要将其移动到性能更低、容量更大的层次上，以保证存储的成本较低。

00:21:27：

你可能也会为了遵循地方法规而移动数据。在欧洲的一些地区接收数据时，数据来源必须保持在特定的政治边界内。

00:21:39：

在某些行业，像 HIPAA 这样的东西限制了数据的移动方式。我认为，随着现代 IT 组织越来越多地分布在不同的数据中心、公有和私有云中，统一地、自动化地管理它们的能力正变得越加重要。

00:21:58 - Saron Yitbarek：

当你想到未来我们要如何管理和存储数据，以及如何处理数据的时候，开源在其中扮演了怎样的角色？你曾提到，你之所以要创建一个开源的解决方案，是因为你个人的理念和你对自由和开源软件的强烈感情。

00:22:16：

你如何看待开源对未来其他解决方案的影响？

00:22:21 - Sage Weil：

我认为，特别是在基础设施领域，解决方案正在向开源靠拢。我认为原因是基础设施领域的成本压力很大，特别是对于构建软件即服务（SaaS）或云服务的人来说，低成本的基础设施很重要，从他们的角度来看，开源显然是一个非常好的方法。

00:22:48：

第二个原因更多地是社会因素，在这个快速发展的领域里有如此多新的工具、新的框架、新的协议、新的数据思维方式，这个领域中有这么多创新和变化，有这么多不同的产品和项目在相互作用，所以很难以传统方式做到这一点，比如说，让不同的公司互相签订合作协议，共同开发。

00:23:20：

开源可以消除此事上的所有阻力。

00:23:28 - Saron Yitbarek：

Sage Weil 是红帽公司的高级咨询工程师，也是 Ceph 项目的负责人。我要绕回到《经济学人》的 Kenneth Cukier，以从一个更整体的视角上进行讨论，因为我希望，我们能够记住他关于人与数据之间关系的看法，以及我们从泥板，到印刷机，再到像 Sage 打造的云端奇迹的进步历程。

00:23:55 - Kenneth Cukier：

这关乎人类的进步，关乎我们如何更好地理解世界，如何从现实中总结经验，以及如何改善世界。这进步也是人类一直以来的使命。

00:24:08 - Saron Yitbarek：

使命永无止境。但是，与此同时，学会处理我们收集到的数据并将其投入使用，是整整一代人的开源任务。我们将在田纳西州的Oak Ridge National Laboratory橡树岭国家实验室短暂停留，结束我们的数据之旅。它是世界上最快的超级计算机 Summit 的所在地，至少在 2018 年是最快的超级计算机。

00:24:43：

这台机器每秒能处理 20 万亿次计算。换个计量单位，就是 200 petaflops。这样的处理速度，对于医院、银行或者今天所有受益于高性能计算的成千上万的组织来说并不现实。

00:25:04：

像 Summit 这样的超级计算机更多的是留给强子对撞机的领域。不过话说回来，我们曾经在泥板上记录的只是 100 字节的信息。

00:25:16：

在数据存储和数据处理的领域中，非凡的壮举不断成为新的常态。有一天，我们或许能将 Summit 级别的超级计算机装进口袋。想一想，到时候我们能够搜索到的答案。

00:25:42：

下一集，我们聊聊无服务器。第 7 集将会讲述我们与基于云的开发之间不断发展的关系。我们将会探究，在我们的工作中有多少可以抽象化的部分，以及在这个过程可能会失去的东西。

00:25:58 - Saron Yitbarek：

同时，如果你想深入了 ChRIS 的故事，请访问 redhat.com/chris ，了解它是如何构建的，以及如何为项目本身做出贡献。

00:26:12 - Saron Yitbarek：

《代码英雄》是一款红帽公司原创的播客。你可以在 Apple Podcast、Google Podcast 或任何你想做的事情上免费收听。

00:26:24 - Saron Yitbarek：

我是 Saron Yitbarek。坚持编程，下期再见。

什么是 LCTT SIG 和 LCTT LCRH SIG

LCTT SIG 是 LCTT Special Interest Group 特别兴趣小组，LCTT SIG 是针对特定领域、特定内容的翻译小组，翻译组成员将遵循 LCTT 流程和规范，参与翻译，并获得相应的奖励。LCRH SIG 是 LCTT 联合红帽（Red Hat）发起的 SIG，当前专注任务是《代码英雄》系列播客的脚本汉化，已有数十位贡献者加入。敬请每周三、周五期待经过我们精心翻译、校对和发布的译文。

欢迎加入 LCRH SIG 一同参与贡献，并领取红帽（Red Hat）和我们联合颁发的专属贡献者证书。

---

via: https://www.redhat.com/en/command-line-heroes/season-2/the-data-explosion

作者：Red Hat 选题：bestony 译者：TimeBear 校对：Northurland, wxy

本文由 LCRH 原创编译，Linux中国 荣誉推出