第 14 章 在云上使用R语言

“如果有人问我什么是云计算，那我一时半会儿还真说不清楚。我会告诉他们，简单地说，云计算就是一种运行业务的更好方式。”

——马克·贝尼奥夫，云计算软件服务提供商赛富时（Salesforce.com）CEO

因为我不是一个想从云计算中获取利润的公司CEO，所以还是给云计算下个定义。我很喜欢微软公司的这个定义（参见https://azure.microsoft.com/en-us/overview/what-is-cloud-computing）：

简单地说，云计算就是通过互联网（云）提供的计算服务——服务器、存储、数据库、网络、软件、分析，等等。提供这些计算服务的公司称为云供应商，通常根据使用情况对云计算服务进行收费，这和你为家里的水电付费非常相似。

如果你还没有使用云服务进行机器学习，那么我可以保证在不远的将来，你会使用的。我知道有些人会担心数据失控、安全问题等，一个创业公司的CEO就问过我这样的问题。我会反问有这种担心的人：“如果你们认为保存在笔记本电脑上的数据是安全的，那么会通过Wi-Fi来访问这些数据吗？”如果答案是“会”，那么这就是在使用云服务，与上面所说的云服务的区别只是存储硬件的环境不一样。

就是这样。你想把办公室装满服务器，弄得像个地牢一样？还是想让别人通过他们安全的、有冗余备份的和遍布全球的基础设施来解决问题呢？

基于云计算使用R语言可以在多个工作地点之间达成无缝连接，也可以获得极大的计算能力，而且可以按照需要快速向上或向下扩展。云计算可以节省大量成本。

有很多种方式可以在云上使用R，我要使用的是亚马逊的云服务平台AWS（Amazon Web Services）和他们的弹性计算云（Elastic Compute Cloud，EC2）。我使用亚马逊云服务进行说明是因为，它是我使用的第一个云服务，而且我已经非常熟练了。这并不是说我认为它比其他产品要好。我现在不这样认为，将来也不会，除非杰夫·贝佐斯（亚马逊集团董事会主席兼CEO）选择我去执行一次载人航天任务，我的态度才会改变。

无论如何，本章的目的是带领并教会你在不写一行Linux代码的情况下，快速地在云上使用R语言和Rstudio。为了最大限度地利用AWS的能力和它花样繁多的在线工具，你可以学习Linux代码，并通过**SSH**（Secure Shell，安全外壳协议）来使用。对于本章的内容，我们要创建并启动一个名为**实例**（instance）的虚拟机，然后通过网页浏览器登录Rstudio，并使用其中的一些功能。网上有很多这样的教程，但我的目标是使你以最简单和最快速的方式开始，并且**今天**就可以在云上使用R。

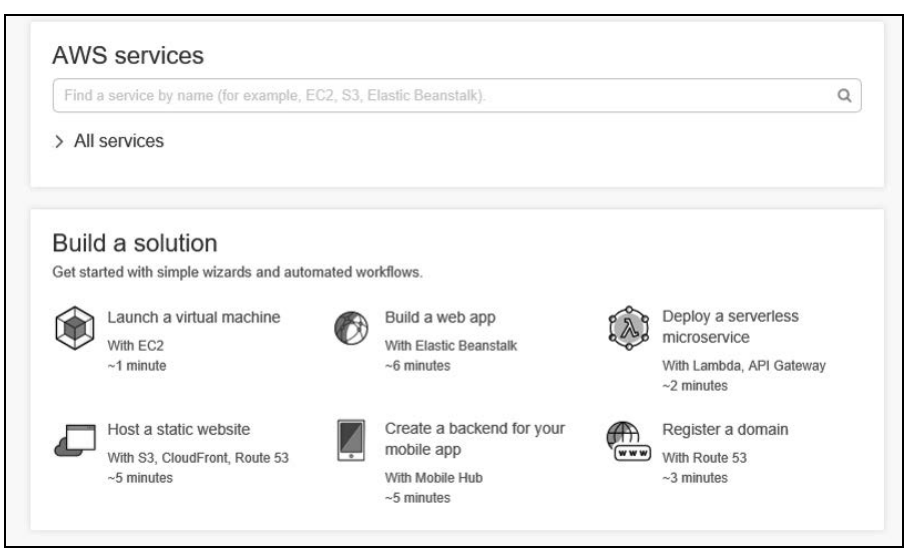
14.1 创建 AWS 账户

第一件事就是要注册一个AWS账户：

https://aws.amzon.com/

这是这个练习的唯一前提条件。注册过程需要一个信用卡，但是你在本章中要做的练习不会花一分钱，因为我们使用的是免费实例。然后你就可以快速启动一个新的实例，这个实例具有你需要的强大计算能力，还可以在完成练习后停止或结束。当你创建账户并登录时，可以选择是否建立安全组。创建实例时，我会通过建立一个新的安全组来说明它的用处。安全组可以让你控制谁可以访问实例，以及如何访问实例。还有，现在先不要创建**密钥对**，除非你特别需要，否则可以在以后创建。

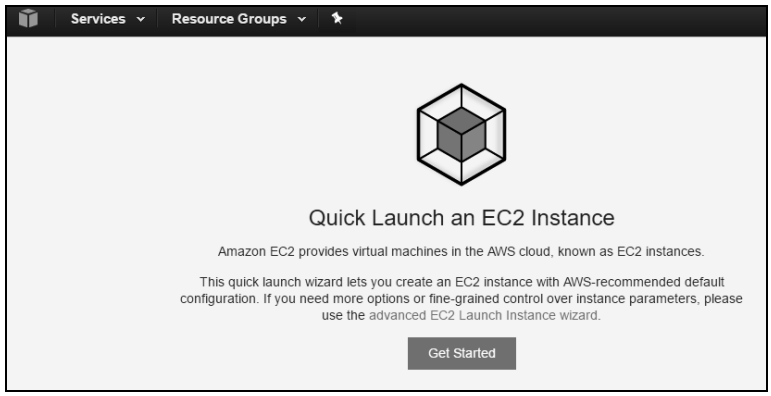
这些步骤都完成之后，登录进入你的AWS控制台，如下所示。



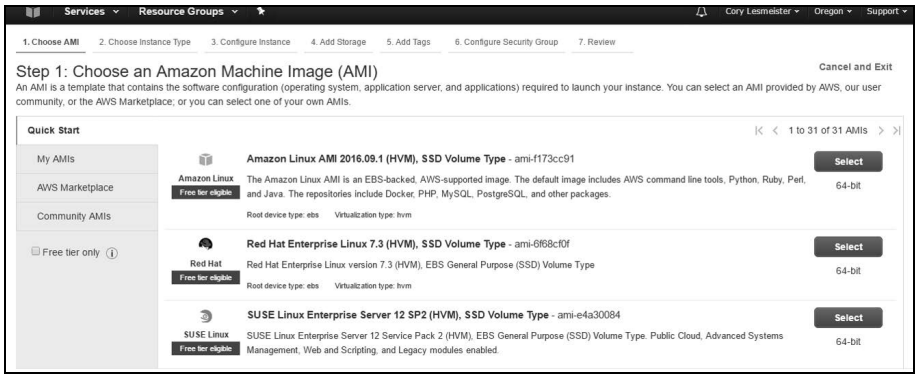
点击智能超链接**Launch a virtual machine**即可创建并启动一个虚拟机。

14.1.1　启动虚拟机

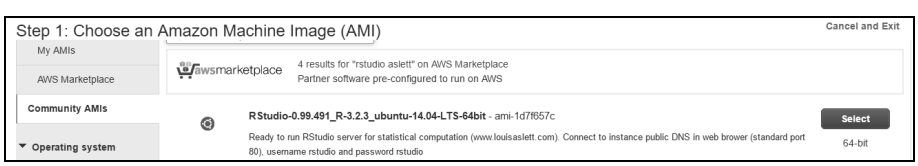
启动虚拟机的超链接会带你进入如下页面。



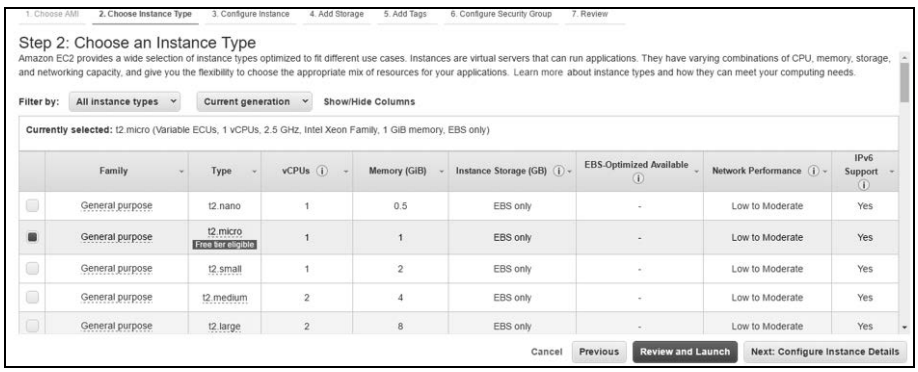
不要点击**Get Started**按钮，而是点击**advanced EC2 Launch Instance Wizard**，如下所示。



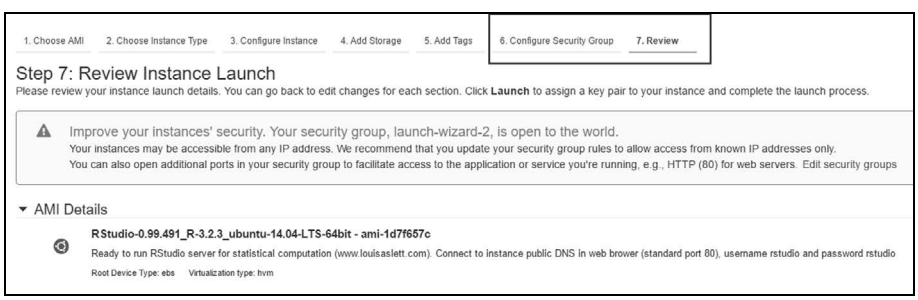
如果你已经有了一些经验，就可以使用不同的**亚马逊机器镜像**（Amazon Machine Image，AMI），并可以定制在AWS上使用R的方式。但本章目标是快速而又简单地在云上使用R，考虑到这一点，AWS用户建立了一些社区AMI，其中已经包含了R和Rstudio。所以，可以在**Quick Start**之后点击**Community AMIs**，这时会弹出一个搜索框，我建议先使用由Louis Aslett维护的AMI：http://www.louisaslett.com/RStudio_AMI/。搜索rstudio aslett可以找到这个AMI，会显示下面的网页，点击Select按钮，如下所示。



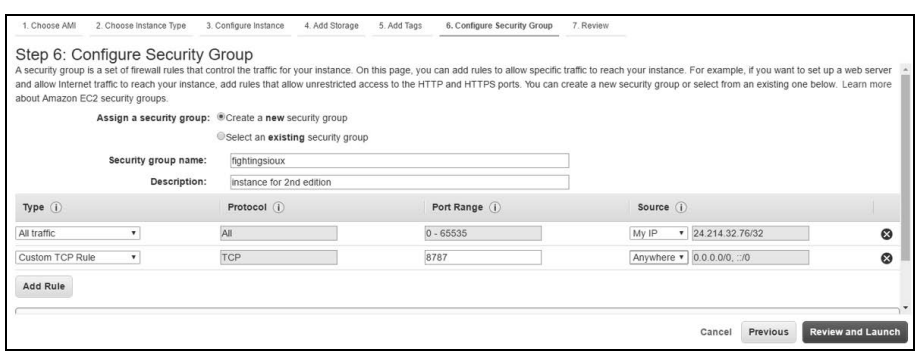
这样就到了第二步——选择实例类型。我选择的是t2.micro免费层实例。



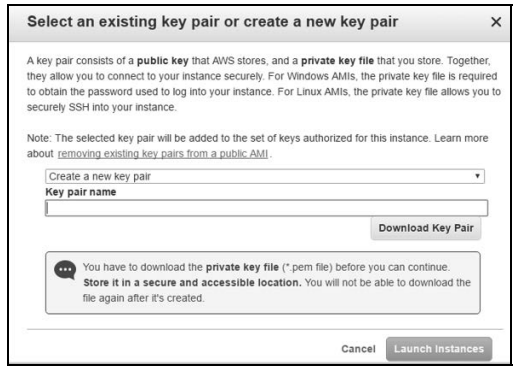
如果已经选择了需要的实例类型，则可以点击Review and Launch。因为这是一个已有的AMI，所以可以跳过第七步——Review标签页。此时可以启动实例，但我们点击一下**第六步——Configur Security Group**。



在启动过程的这一步，你可以建立一个安全组，也可以使用一个现有的。下面是创建**新安全组**的示例。



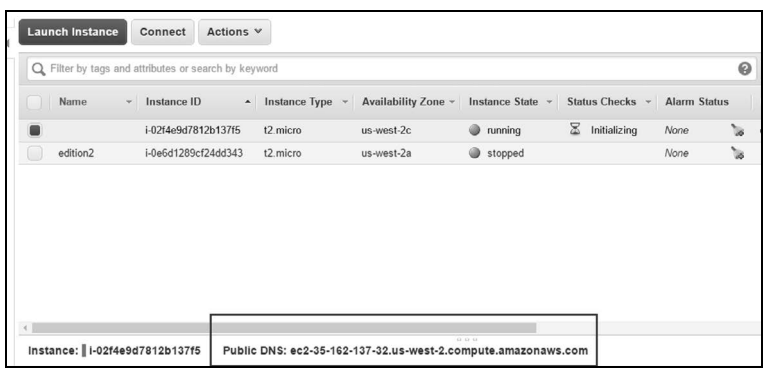
完成该步骤后（可以完全不做修改），点击**Review and Launch**。这会带你回到**第七步**，此时可以点击**Launch**，来到选择新密钥对或现有密钥对的页面。



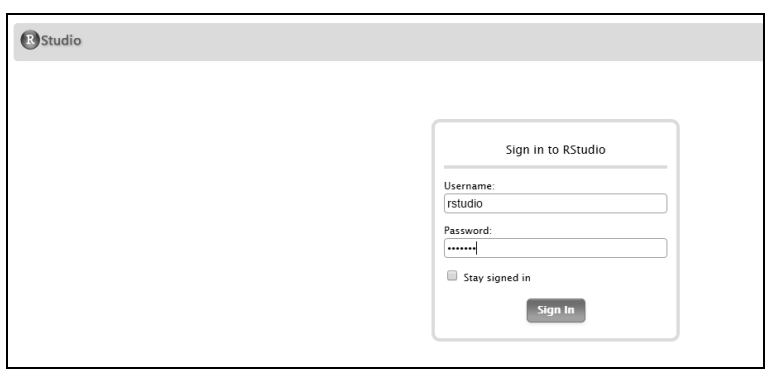
完成之后，点击**Launch Instance**，回到AWS控制台。

14.1.2　启动 Rstudio

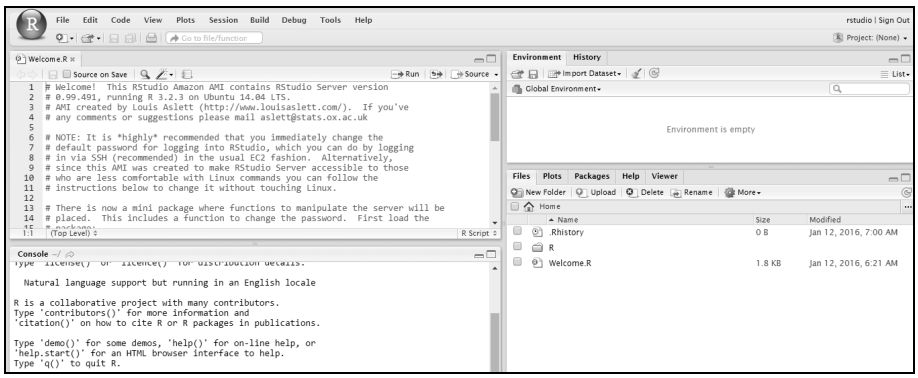
启动实例之后，回到AWS控制台并选择该实例时，你会看到如下页面。



注意选中实例的**Public DNS**。要在你的网页浏览器中启动Rstudio，有这个就足够了。要在浏览器中启动Rstudio，需要先来到Rstudio的登录页面，用户名和密码都是rstudio。



就是这样！你已经在一个虚拟机上运行Rstudio了。它看上去应该如下所示。



在左上角的**Source Panel**窗格中，有关于如何修改密码和连接到Dropbox的指示。

为了演示如何从网络上加载数据，我要从github上加载一个前面章节中使用过的.csv文件。试试climate.csv好吗？首先要安装和加载RCurl包：

```
> install.packages("RCurl")

> library(RCurl)
```

然后，需要获得GitHub上的数据连接：

```
> url <-
"https://raw.githubusercontent.com/datameister66/data/master/climate.csv"
```

然后，将数据读取到Rstudio中：

```
> climate <-read.csv(text = getURL(url))
```

确定读取数据的结果：

```
> head(climate)
  Year CO2   Temp
1 1919 806 -0.272
2 1920 932 -0.241
3 1921 803 -0.187
4 1922 845 -0.301
5 1923 970 -0.272
6 1924 963 -0.292
```

就是这样。现在，你已经成为一个基于云计算的机器学习勇士了，你几乎可以像使用自己的计算机一样在虚拟机上进行各种操作。

请注意，如果你完成操作并退出Rstudio，请一定回到控制台停止实例。

14.2　小结

在本书最后一章中，我们介绍了如何快速而又简单地在云上运行R语言和Rstudio。在这个练习中，我们通过使用AWS，循序渐进地介绍了如何在云上创建虚拟机（实例）、配置虚拟机、启动虚拟机，以及在浏览器中运行Rstudio。最后，通过从GitHub上读取climate.csv文件，说明了从网络上加载数据有多么容易。通过本章对云计算的简单介绍，你可以在任何有互联网连接的地方开展工作，并且可以对实例进行快速扩展和收缩，以满足你的需求。本书的主要章节到此结束。我希望你能喜欢本书内容，并能将书中介绍的方法和你逐渐学会的其他方法在实际中加以应用。谢谢！