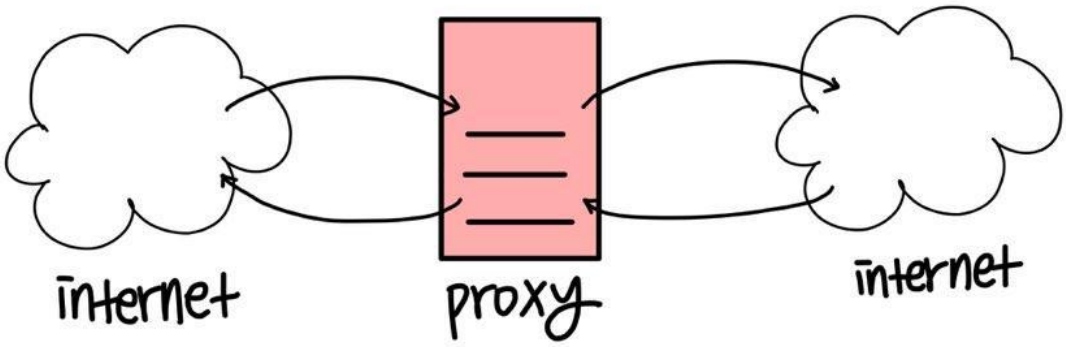
爬虫对代理的需求

代理服务器是一种服务器程序或设备，用于调解从寻找资源的客户端到在计算机网络中提供这些服务的服务器的请求。

由于爬虫需要从单个 IP 地址向服务器发出大量请求，服务器可能会识别过多的请求并阻止 IP 地址以防止进一步抓取。 使用代理服务器来防止阻塞，即使IP地址发生变化，抓取也将继续正常运行。由于它创建了匿名性，它还有助于掩盖机器的 IP 地址。



代理服务器到底是什么？

在最终用户和互联网/网站之间，代理充当中间人。 从本质上讲，它是一个门户网站，用户可以利用它来浏览网站而无需透露他们的个人 IP 地址。

IP地址是一个特殊的地址，当用户通过它连接到互联网时，计算机使用它来识别自己。 使用代理服务器时，连接不是直接连接到互联网，而是通过代理服务器重定向，代理服务器在使用自己的 IP 地址代表您联系网站之前控制流量和请求。

互联网安全、互联网流量的负载平衡或隐私问题是使用代理的三个主要原因。 现在您知道代理是什么，让我们探讨一下这个工具对于爬虫的重要性。

为什么爬虫需要代理？

从网站管理员的角度来看，从单个 IP 地址重复向网站发送流量似乎是一种攻击。 因此，网站将始终制定政策来阻止、限制或将被视为攻击它们的 IP 地址列入黑名单。 控制这种在线抓取流量的最简单方法是通过代理。 为了分散请求并匿名抓取，可能会使用代理或免费代理。

代理服务器对于小规模的爬虫并不是特别必要。 但是，如果您的爬虫需求更复杂，例如从某个区域获取数据或大量抓取，则代理是必需的。

有哪些类型的代理？

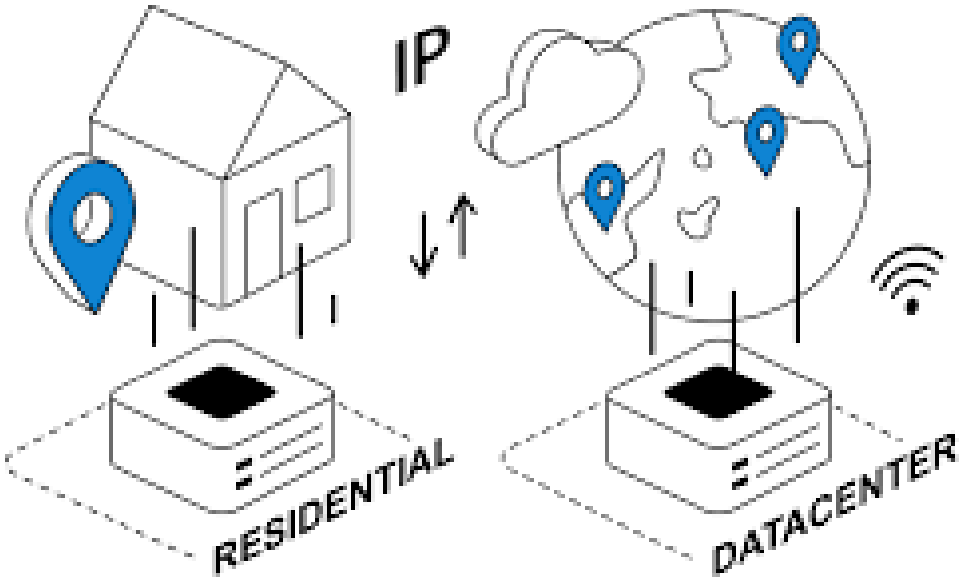
爬虫有多种代理类型，具体取决于用例。

静态代理：

最基本的代理服务器就是那些。 尽管最容易检测，但它们价格低廉、快速且可靠

住宅代理:

这些使用实际用户的设备和各种各样的 IP 地址。 它们极难检测、昂贵、缓慢且不可靠，因为用户可能会关闭他们的小工具或失去互联网连接。



专业代理：

这些用于特定用途，例如社交媒体网站或 Google 搜索结果页面。

移动代理：

这些使用实际移动设备的 IP 地址。 网站更倾向于信任移动设备，因为用户很可能是人。

代理用于网页抓取有什么优势？

查看在爬虫时使用2808代理服务器解决方案的一些典型优势。

隐私浏览

鉴于爬虫的性质，您可能不想透露您的设备标识。 您的私人 IP 地址可能会被追踪，您可能会成为广告的目标，或者如果网站识别您的身份，您甚至可能无法访问该网站。 通过使用代理，您可以使用代理服务器的 IP 地址而不是您自己的 IP 地址。

防止 IP 封锁和禁令

使用代理的另一个好处是它可以防止您的 IP 被列入黑名单。 大多数现代网站都启用了抓取数据限制和其他反机器人检测机制。 这些限制了 scarper 可以向站点发出的请求的数量。 但是，您可以通过使用代理池通过多个不同的 IP 地址传送流量来规避速率限制等问题。

访问基于位置的信息

一些网站禁止来自不同国家的访客。 根据您的 IP 地址所在的位置，它们会启用特定区域的内容并仅显示特定内容。 您可以通过在必要位置使用代理来访问该内容。 以多种货币获取价格信息是电子商务中常见的例子。

协助重大项目抓取

对于从网站收集数据所需的时间至关重要的大容量抓取项目，使用代理是抓取网站的最佳实践方法。您可以执行并行会话并通过使用大量代理来加快数据抓取的速度。

住宅代理

如果您从一个 IP 地址发送过多的 HTTP/HTTPS 请求，网站管理员无疑会限制您的 IP 地址，以防止进一步的数据挖掘。 理想的选择是构建一个代理池，并在来自单个代理服务器的预定数量的请求之后轮换或迭代成员。