



爬虫程序管理的 10 个首要考量因素

电子书



目录

简介	03
1. 复杂专业知识	04
2. 情报	05
3. 弹性保护	06
4. 误报与漏报	07
5. 灵活的对策	08
6. 部署	09
7. 监测和报告	10
8. 保护 API	11
9. 网站还是网页	12
10. 托管服务	13

简介

想知道爬虫程序问题变得有多么严峻吗？您试着抢购一次泰勒·斯威夫特的演唱会门票或新款 AJ 运动鞋就知道了。而这还只是一些炒作营销活动而已。爬虫程序在各个行业中都变得愈加普遍且充满危害。

对于想要寻求破解办法的人来说，更加糟糕的是爬虫程序管理规则已发生变化。说实话，它一直都在不断变化。爬虫程序管理往往被描述为一场军备竞赛或者一场猫鼠游戏，各个企业在不断地建立防线，而爬虫程序创造者在不断地寻找攻破这些防线的方法。但是，现在不断发展的并非只有爬虫程序自身。它们所处的环境也在不断变化。举例来说，企业要应对的不再只是单个攻击者，还有团伙配合攻击。而且，现在可以租用某个爬虫程序一个星期，就像您在爱彼迎上租个房子那样简单。同样地，解决方案也无法再简单地将爬虫程序分为良性和恶意爬虫程序两个类别。因为现在有更多的灰色地带。

爬虫程序及其所处环境的这种发展变化让如何选择爬虫程序管理软件变得比以往更加困难。您不仅需要确定哪些措施对以往的爬虫程序有效，还需要确定哪些措施对现在和未来的爬虫程序有效。

本指南概述了一些关键考量因素，供买家在选择爬虫程序管理软件时参考。
阅读本指南，理清并避开一些可能影响您做出明智决策的干扰因素。

1 复杂专业知识

顾名思义，爬虫程序管理解决方案的任务就是检测爬虫程序。换言之，它们会寻找自动化痕迹以及表明请求者不是人类的其他迹象。但是，随着爬虫程序发生演变并变得更加复杂，爬虫程序也变得更加专门化。现在，爬虫程序的设计目的很明确，比如抓取您网站的内容，在您的营销活动中抢购，以及实施撞库攻击劫持客户的帐户。而且，针对某一类专业爬虫程序的检测方法无法检测其他爬虫程序。您需要知道供应商能否阻止您所面对的特定爬虫程序，而不仅仅是基本的通用爬虫程序。

考量因素：

- 供应商是否有基于业务应用场景的专业爬虫程序检测方法？
- 供应商能否展现相关专业知识，以解决您所面临的具体爬虫程序问题？
- 供应商的其他客户中有多少面临相同的问题？供应商从这些客户身上学到的经验教训是否会让您受益？
- 提供商是否提供报告、服务或其他功能来进一步支持您抵御专业的对抗性爬虫程序？



2 情报

爬虫程序管理解决方案的优劣取决于能不能识别所监控的爬虫程序的特征。虽然一些供应商可能宣称能检测出 99.9% 的爬虫程序，却无法客观地衡量其有效性。爬虫程序始终在发生变化，因此您以前捕捉到的爬虫程序现在很可能已学会如何规避这种检测方法。在爬虫程序管理工具时，有一种更好的标准就是看看供应商是如何说明自己的爬虫程序检测能力的。您需要的是既能够检测极为复杂的爬虫程序（而不是常见的爬虫程序），又能够从大型数据集中提取数据的解决方案。请谨记，很多人工智能 (AI) 和机器学习 (ML) 工具都是开源的，因此在评估某个解决方案的 AI/ML 时，该解决方案向算法提供的数据量、数据清洁度以及提供数据的速度都是未得到重视的考量因素。相关见解应包含多个域中每次登录时的信任指标和风险评估。此外，有效的解决方案应使用最新方法，采取多管齐下的手段来进行爬虫程序检测。

考量因素：

- 请求供应商详细说明自己的爬虫程序检测能力。如果供应商有大量对攻击者有吸引力的客户，他们将拥有更多经验以及更全面的数据集，在这些数据集的基础上可以构建众多功能，包括要评估的风险和信任信号、更多设备异常检测等。应该将缺乏透明度当作是一种警告信号。
- 供应商是否使用 AI/ML 为解决方案提供支持？这些模型的复杂程度如何？同样重要的是：供应商向这些模型提供了多少数据？攻击者肯定在使用 AI。您也应该使用。
- 但是，使用 AI 还远远不够。供应商是否拥有一支由安全研究人员和威胁情报分析师等合格专家组成的团队在不断寻找新的攻击技术并监控黑客社区以确保您始终能够先发制人？

3 弹性保护

在您拦截复杂的爬虫程序时，它并不会永久消失。它随时都会回来，而且始终都在发生变异，企图躲避您的检测。很多爬虫程序管理解决方案最开始都能检测出爬虫程序（或者至少其中的一部分），但随着爬虫程序开始发生变异，它们便不再有效。据 Akamai 观察，爬虫程序在数小时内就会发生变异。传统开发周期太过漫长而无法跟上节奏。确保您选择的解决方案能够随着时间推移不断学习和发展，最好能够自动使用机器学习。这应当包括先发制人的防御措施。对于寻求获取相关信息以用于规避防御机制的攻击者来说，这些措施能够增加他们的攻击难度。

考量因素：

- 寻找采用先进爬虫程序检测技术（如用户行为分析和特定于客户的学习模型）的解决方案。这些解决方案将随着爬虫程序的变异在更长时间里保持效力。
- 确定该解决方案是否包含 JavaScript Obfuscation 技术等防御性策略，该技术能够让攻击者更难以对能够偷偷绕过防御机制的爬虫程序进行反向工程。
- 请已部署该解决方案的其他客户提供证明或参考，了解其能否长时间保持效力。





4 误报与漏报

当爬虫程序管理解决方案显示它拦截了爬虫程序时，如何才能知道该系统拦截的不是合法用户？很多供应商出手很快，对误报相当宽松。如果他们缺少能够根据每次检测结果对爬虫程序进行评分的解决方案，则他们可能无法检测灰色爬虫程序，从而导致产生二元“是/否”决策。他们往往喜欢向客户表明自己拦截了很多“爬虫程序”，哪怕误报率其实很高，也就是说他们在阻止爬虫程序的同时也在阻止有效流量，即对贵公司有价值的人或“良性”爬虫程序。另一方面，低误报率听起来很不错，但供应商的以下做法却难以让人接受：供应商为此不得不降低解决方案整体有效性来确保它未拦截人类用户，最终让原本应该拦截的爬虫程序畅行无阻。您希望在不妨碍正常业务的同时拦截恶意爬虫程序，但是也不希望降低防护能力。您需要放心地知道与您合作的供应商也非常关心准确性以及误报和漏报的影响。

考量因素：

- 供应商是将调整误报/漏报的任务留给您处理，还是会在功能与服务上进行投入以便与您协同工作？
- 该解决方案是否会从整个网站的流量模式中进行学习并自动进行自身优化以减少您团队的负担？
- 供应商是否建议使用验证码来代替其他质询操作？这通常将一家供应商的无能暴露无疑。虽然用户不喜欢验证码，但是，供应商提供验证码要比调整规则以最大限度降低误报来得容易。
- 您能否察明该解决方案为什么将某个请求标记为来自爬虫程序？解决方案是否根本就是一个黑箱？寻求一种解决方案，以能够通过精细地监测请求来验证所采取的对策，以及在更改设置时，能够先监测这些更改可能带来的影响，然后再应用到生产环境之中。

5 灵活的对策

您会很容易地认为自己只需要考虑如何在允许良性爬虫程序进入的同时拦截恶意爬虫程序即可。但环境已变得远比这种情况复杂。很多爬虫程序操纵者学会了如何将风险信号降低到足以让其爬虫程序处于灰色地带，因为他们知道大多数企业宁可冒险让恶意爬虫程序进入也不愿冒险拦截合法用户。您的解决方案应当提供一组复杂的对策，不仅能实施拦截或允许，还包含加密质询和逐步 MFA 等质询操作。另外，解决方案还应当包含用于处理其他类型的情况（如良性爬虫程序）的对策。您可能希望在高流量时段降低合作伙伴爬虫程序的速度，并在非高峰时段让这些爬虫程序立即进入。另外，您还可以选择对同一已知类别中的爬虫程序采取不同的对策。例如，如果您是零售商，那么您可能会允许更受欢迎的优惠券爬虫程序检查您的网站，同时拦截不想与其进行业务往来的爬虫程序。您需要根据爬虫程序对业务和 IT 的影响灵活地对不同类型的爬虫程序采取不同对策，特别是它们的影响会随位置、一天中的时段或季节性而发生变化时。不仅如此，您的解决方案还需要符合以下条件：既不会简单地拦截所有爬虫程序（并且这样做会教会爬虫程序更改入侵策略），又能够制造障碍，让攻击者面临更大的困难和更高的成本。

考量因素：

- 该解决方案能让您为不同类型的爬虫程序创建不同的类别，还是只区分好坏？
此外，它能否针对同一类别中的爬虫程序（如搜索引擎或金融聚合器）创建不同的对策？
- 该解决方案支持哪些类型的条件对策？它是否支持有助于更好地调整流量的高级对策（如限速和提供替代内容）？它是否包含加密质询等操作？
- 该解决方案对于管理您见到的不同爬虫程序有多灵活？是不加区别的反复操作，还是能够根据一天的不同时段、按流量百分比或按 URL 精准地采取对策？
- 除了显示醒目的 403 错误之外，该解决方案能否引入一些资源密集型问题，让爬虫程序操纵者面临更高成本，同时降低高请求量攻击的速度？



6 部署

谈到任何爬虫程序管理解决方案时，需要花费多长时间启动该解决方案以及您能够在多短的时间内对其进行修改应当成为核心考量因素。买家应当对需要修改其现有应用程序或影响应用程序性能的任何解决方案保持警惕。部署延迟会带来高昂的成本，并且如果您需要根据业务活动（如限时抢购）的需求更改应用程序，这只会需要更多资源。

考量因素：

- 该解决方案是否会实时运行，同时不会影响现有应用程序的性能？
- 它是否需要您更改现有应用程序？
- 它能否进行纵向扩缩以应对不可预见的事件（如容量耗尽型攻击）或预期活动（如限时抢购）？



7 监测和报告

每种爬虫程序管理解决方案都能为您显示对爬虫程序流量的概要统计数据，但您需要的不止是这些。在进行基础架构规划或向您的管理层上报时，尤其适合使用概要统计数据，但这些数据无法提供分析爬虫程序流量所需要的详细信息。同样也无法提供所需的证据，让您可以确信该解决方案采取了正确的对策。对于能拦截用户的解决方案来说，您不想要一个黑箱。您需要详细的报告来为业务提供支持并帮助您更好地了解更改风险阈值会如何影响性能。

考量因素：

- 该解决方案是否提供一定的报告功能，让您能够仔细审视特定的爬虫程序、僵尸网络或爬虫程序特征？它能否报告不同的评分分段、哪些爬虫程序正在攻击哪些端点以及显示已采取哪些措施？
- 您能否调查流量激增现象并查看个别请求？有时您需要查看请求详细信息才能知道怎么办。
- 该解决方案能否显示您的爬虫程序流量与业内其他公司的对比情况？
- 报告机制与其他安全解决方案的报告机制有什么样的关联？您能否全面分析流量，或者是否有单独的视图？



8 保护 API

无论供应商还是解决方案，现在更先进的爬虫程序检测技术都依赖于注入 JavaScript 代码并分析客户端响应。但是，如果基于 API 的客户端对 JavaScript 没有响应，您要怎么处理您的 API？如果您需要暴露 API 来支持移动应用程序或其他第三方，那么您需要能帮助您像保护网页一样保护 API 的解决方案。否则，您的爬虫程序（和爬虫程序问题）只是从网页迁移到 API 而已。

考量因素：

- 供应商为 API 提供何种保护？只是定额管理和速率限制吗？
- 寻找能将供应商的先进爬虫程序检测融入您的移动应用程序的移动功能。
- 尽管无法始终像其他主动检测一样有效，但基于声誉的方法仍不失为一个好的选择，它可以保护为那些可能无权访问 SDK 等移动功能的第三方提供支持的 API。



9 网站还是网页

如果您的网站不止一个页面，您可能会遭遇多个爬虫程序问题，每个问题影响网站的不同部分。价格抓取会对您的产品页面带来很大影响。内容抓取会危害您的增值数字内容。与此同时，针对您登录页面的撞库攻击仍在不断发生。但是，对于爬虫程序管理解决方案来说，一些解决方案旨在只解决一个问题。确保您的管理解决方案能帮助您解决所有爬虫程序问题，无论这些问题影响整个网站，还是只影响特定网页。

考量因素：

- 该解决方案的关注点是什么 — 个别网页还是整个网站？
如何部署 — 在个别网页还是整个网站的前面？
- 该解决方案能否帮助您解决所有爬虫程序问题，无论是撞库攻击、网络内容抓取还是内容聚合？





10 托管服务

您需要管理爬虫程序，以控制其对您和您业务的影响，但是爬虫程序管理并非易事。虽然您的公司可能具备相应的专业知识，但有时候仍然需要额外的帮助，让精通爬虫程序问题的专家来帮助您。此外，为这些职位配备相关人员变得愈加困难。如果有人离职，会发生什么？任何人都能查看 HTTP 请求并创建签名来阻挡流量，但是，这并不能解决您的问题。您需要的是有人能将爬虫程序与您的核心问题联系起来，设计并实施一种策略来解决这些问题。

考量因素：

- 公司内部是否具备特定于爬虫程序的必要专业知识，让您可以利用任何解决方案？
- 爬虫程序管理供应商提供专业服务还是只销售产品？
- 在发生紧急情况时，您是否具备随时可用的主动监控和补充性专家资源？





主动应对，而不是被动等待

在爬虫程序成为问题，并且下一波发展变化让现有防御措施形同虚设之前，投资于爬虫程序管理会更好地为您的服务。在研究如何选择相关产品时，请考虑这些因素。Akamai Bot Manager 可为您提供所需的保障。如需了解详情，可申请获取包含模拟攻击的个性化导览。

了解更多



扫码关注，获取最新CDN前沿资讯

Akamai 可在您创建的一切内容和体验中融入安全性——无论您在何处进行构建，将其分发到何处，从而保护您的客户体验、员工、系统和数据。我们的平台具备监测全球威胁的能力，这让我们得以帮您调整并增强安全状况，从而实现 Zero Trust、阻止勒索软件、保护应用程序和 API 或抵御 DDoS 攻击，让您信心十足地不断创新、发展和转型。如需详细了解 Akamai 的云计算、安全和内容交付解决方案，请访问 akamai.com 和 akamai.com/blog，或者扫描下方二维码，关注我们的微信公众号。发布时间：2023 年 9 月