webhose.io

Web Crawling - Build or Buy?

So, you need to crawl the web and download structured content for your software application. Should you run it yourself or use the Webhose.io solution?

Option 1: Build

Crawler Software

You can either write a proprietary solution that will be tailored to your requirements, or modify an existing open source crawler to match your specific needs. Either way, this initial setup should take a few months to develop. If you want to add new sites and make sure your crawler downloads the relevant content from each source, you will need at least one dedicated developer to either build or maintain the crawler. The costs of hiring a developer ranges, but the minimum should be around $3,000 a month. Also if you already have dedicated resources, setting up and maintaining your crawlers and the software will certainly consume a lot of their valuable time.

Hosting

You need to run your crawler 24/7 on an online server. You can get a basic dedicated machine for about $100 a month that would be enough to run a multi-threaded crawler, that can download thousands of pages a day from multiple sites. It really depends how efficient your crawler software is

Scaling

Content is flowing in, things are moving forward - great! Now you want to scale up, add more crawlers, filter the content you crawl, maintain hundreds or thousands of sites. You now need a database to filter out duplicates and to schedule crawling jobs between multiple crawlers. If you want to filter the data, you need an indexing solution to query against. You need to hire more developers to code a more complex solution and to maintain the additional sites you add as sites are dynamic and constantly change. Rates are climbing through the roof.

Option 2: Buy

Here at Webhose.io we have already built a server farm with thousands of crawlers, working 24/7 to download millions of web pages daily. We crawl millions of sources, our dedicated team knows how to maintain and of course very efficiently add new sources. We remove duplicates, and enable you to filter only the content you require for a fraction of the cost of running a crawling operation yourself.

When you use Webhose.io, you know that you are using a best-of-breed DaaS (Data-as-a-Service) solution, used by enterprise organizations like Salesforce Radian6, Meltwater, Sysomos, Engagor, Kantar Media and many others. There is no second guessing, Webhose.io provides you with more coverage, frees up your resources so that they can be used elsewhere, gives you the ability to quickly and easily add new sources as and when required, and of course saves you a lot of money!

Conclusion

	Bare minimum in-house solution yearly cost	Webhose.io yearly cost
Initial setup (estimated 3 month work)	$9,000	-
Hosting	$1,200	-
Ongoing Development & Maintenance	$36,000	-
Monthly Subscription	-	$200
Total	**$46,200**	**$2,400**

You save a whopping 95% or $43,800 a year and can filter the data from all the sources on our crawling list. Of course the more data you need the more you save.

It really is a no brainer. Free up your resources and let the experts collect and provide you with the data in a structured format so that you can focus less on data sourcing and more on data analysis. You will instantly have access to more high quality data from hundreds of thousands of sources and of course all this while reducing your costs substantially.

By the way, if you can find a good developer for $3,000 per month please let us know as we would be happy to speak with him / her :)

Webhose.io

Webhose.io works with many organizations from different industries both big and small providing a best-of-breed Data-as-a-Service solution. To learn more about Webhose.io please visit us at our website webhose.io

To see how Webhose.io can help your organization with higher quality data from more sources please contact:

David Sassoon
david.s@webhose.io
Tell. +972-54-8061266
webhose.io