Log Line Contents for OnCrawl SEO Log Analysis

Make sure your log lines contain all the information you need.

REQUIRED FOR LOG ANALYSIS

❑ **Query path** (/blog/) or full URL (https://www.oncrawl.com/blog/)

WHY? This indicates which page or resource the bot or the user wanted to view. We use this to calculate the number of hits per page, to filter or count hits by page group, to cross-analyze data for a given URL, and much more.

❑ **Date and time**

WHY? This indicates when the page or resource was requested. We use this to establish crawl frequency, to create time graphs, to allow you to filter by date, and to target data for the right time period for an analysis.

❑ **User Agent**

WHY? The User-Agent contains the essential information about who is making the request: type of device, and--more importantly--the bot name if the visitor is a bot.

❑ **Status code**

WHY? The server returns a status code for every requested item. We use this to establish inconsistencies in status codes reported to users, to OnCrawl, and to Googlebot.

REQUIRED IF YOUR SITE USES HTTPS

Either:

❑ **Scheme** (http or https), if not already present in the full URL

Or:

❑ **Port** of the request (80 on HTTP / 443 on HTTPS)

WHY? This information is necessary to know whether the visitor requested the HTTP version or the HTTPS version of a page.

REQUIRED IF YOUR LOG FILES CONTAIN INFORMATION FOR MULTIPLE SITES OR SUBDOMAINS

❑ **vhost**, if not already present in the full URL

WHY? This information is necessary to distinguish between URLs on one subdomain and URLs on another.

OPTIONAL BUT HIGHLY RECOMMENDED

❑ Referer

WHY? This information passed on by the browser indicates the page from which the visitor is coming.

The referer is required if you intend to use log analysis to identify and examine organic traffic (users coming from a Google search page).

❑ Client IP

WHY? We use this information to confirm that visitors with a Googlebot user-agent are really Googlebots by reverse lookup. This allows OnCrawl to filter out spam bots masquerading as Google. Once the check is performed, we do not keep or save this information anywhere. We do not need or use IP addresses for lines that do not have a Google user agent.

❑ Response size

WHY? This indicates how much data the server transferred for the requested page or resource. It is extremely helpful when identifying page size and when searching for server errors such as pages that have a 200 status but 0 bytes of content.

❑ Response time

WHY? This is the amount of time it took the server to provide the requested page or resource. This helps contribute to an accurate measure of page speed across your site.