Basic Info Points -----

* A web crawler bot, also known as a spider or search engine bot, downloads and indexes content from the Internet.
* Web crawlers are operated by search engines to learn about webpages and provide relevant search results.
* They access websites automatically and collect data.
* Web crawlers start with known webpages and follow hyperlinks to other pages to index content.
* The extent of web crawling on the Internet is estimated to cover only 40-70% of webpages.
* Search indexing creates a catalog for the Internet to retrieve information when users search.
* Indexing focuses on page text and metadata, such as titles and descriptions.
* Web crawlers select pages to crawl based on factors like linking, visitor traffic, and importance.
* They revisit webpages to update content, and robots.txt files control access to web properties.
* Allowing web crawlers access to web properties depends on factors like server resources and privacy.
* Web crawling is different from web scraping, which is downloading website content without permission.
* Web crawlers play a crucial role in SEO by indexing web content for search engine results.
* Major web crawlers include Googlebot, Bingbot, DuckDuckBot, and Yahoo! Search's Slurp.

Top 10s -----

“keycdn” - [link](https://www.keycdn.com/blog/web-crawlers)

Existing Software Examples -----

Google - [Link](https://developers.google.com/search/docs/crawling-indexing/overview-google-crawlers)

"Crawler" (sometimes also called a "robot" or "spider") is a generic term for any program that is used to automatically discover and scan websites by following links from one web page to another. Google's main crawler is called Googlebot.

* The **user agent token** is used in the User-agent: line in robots.txt to match a crawler type when writing crawl rules for your site. Some crawlers have more than one token, as shown in the table; you need to match only one crawler token for a rule to apply. This list is not complete, but covers most crawlers you might see on your website
* The **full user agent string** is a full description of the crawler, and appears in the HTTP request and your web logs.
* **special-case crawlers** are used by specific products where there's an agreement between the crawled site and the product about the crawl process. For example, AdsBot ignores the global robots.txt user agent (\*) with the ad publisher's permission. The special-case crawlers may ignore robots.txt rules and so they operate from a different IP range than the common crawlers. The IP ranges are published in the [special-crawlers.json](https://developers.google.com/static/search/apis/ipranges/special-crawlers.json) object.
* User-triggered fetchers are triggered by users to perform a product specific function. For example, [Google Site Verifier](https://support.google.com/webmasters/answer/9008080) acts on a user's request. Because the fetch was requested by a user, these fetchers generally ignore robots.txt rules. The IP ranges the user-triggered fetchers use are published in the [user-triggered-fetchers.json](https://developers.google.com/static/search/apis/ipranges/user-triggered-fetchers.json) object.

Apache Nutch -

Apache Nutch is an **open-source web crawler** written in Java. It is released under the Apache License and is managed by the Apache Software Foundation. Nutch can run on a single machine, but it is more commonly used in a distributed environment. In fact, Nutch was designed from the ground up to be scalable and easily extensible.