CLAIMS

1. A computer-implemented method for executing a plurality of web scraping requests over time, comprising:

(a) during a first time period, receiving a plurality of web scraping requests from a client computing device, each of the plurality of web scraping requests specifying a target website to capture content from;

(b) placing job descriptions for the plurality of web scraping requests on a queue, the job descriptions each describing at least one web scraping job specified by a request from the plurality of requests;

(c) incrementally over a second time period longer than the first time period, removing the job descriptions from the queue;

when each respective job description is removed from the queue:

(d) generating a web request for the target website as specified in the job description; and

(e) transmitting the web request such that the web request reaches the target website via a proxy selected from a group of proxies.

2. The method of claim 1, wherein the removing (c) occurs in increments selected to avoid the group of proxies from becoming blocked by the target website.

3. The method of claim 1, further comprising:

(f) in response to the web request, receiving content transmitted from the target website; and

(g) transmitting the received content to the client computing device.

4. The method ofclaim 1, wherein the plurality of web scraping requests each include a client identifier indicating a client from which the web scraping request originates, further comprising:

(f) selecting the queue from a plurality of queues based on the client identifier.

5. The method of claim 4, wherein the plurality of web scraping requests each further include an identifier of the target website, further comprising:

(g) selecting the queue from the plurality of queues based on the identifier of the target website.

6. The method of claim 5, further comprising selecting the queue from the plurality of queues based on whether the job description specifies that web scraping be made from a headless browser.

7. The method of claim 6, further comprising, when the job description requests that web scraping be made from the headless browser:

(h) in response to the web request, receiving content specifying a web page, the content transmitted from the target website;

(i) analyzing the content to determine web addresses for additional content needed to render the web page; and

(j) requesting the additional content from the web addresses.

8. The method of claim 1, wherein incrementally removing (c) comprises repeated removing a job description from the queue at a time window.

9. The method of claim 8, wherein incrementally removing (c) comprises repeated removing a plurality of job descriptions from the queue at the time window.

10. The method of claim 9, wherein an amount of job descriptions in the plurality of job descriptions are specified based on a client that sent the plurality of web scraping requests.

11. The method of claim 1, further comprising:

(f) determining that web scraping requests from the client computing device no longer need to be throttled;

(g) determining when the queue is empty; and

(h) when the queue is determined to be empty, deleting the queue.

12. The method of claim 1, , further comprising:

(f) determining which of the plurality of web scraping requests are synchronous requests and which are asynchronous requests; and

(g) prioritizing web scraping requests determined in (f) to be synchronous requests over web scraping requests determined in (f) to be asynchronous requests to help ensure that results from synchronous requests are returned before a connection with the client computing device times out.

13. A system for executing a plurality of web scraping requests over time, comprising:

at least one processor;

a memory configured to store a queue;

a request intake manager configured to, during a first time period, receive a plurality of web scraping requests from a client computing device, each of the plurality of web scraping requests specifying a target website to capture content from;

a throttler configured to place job descriptions for the plurality of web scraping requests on the queue, the job descriptions each describing at least one web scraping job specified by a request from the plurality of requests;

a job remover configured to, incrementally over a second time period longer than the first time period, remove the job descriptions from the queue;

a scraper configured to, when each respective job description is removed from the queue, generate a web request for the target website as specified in the job description, and transmitting the web request such that the web request reaches the target website via a proxy selected from a group of proxies.

14. The system of claim 13, wherein the job remover is configured to remove in increments selected to avoid the group of proxies from becoming blocked by the target website.

15. The system of claim 13, wherein the scraper is configured to, in response to the web request, receive content transmitted from the target website, and the request intake manager is configured to notify the client that the received content is available for transmission to the client computing device;

16. The system of claim 13, further comprising a memory configured to store a plurality of queues, wherein the plurality of web scraping requests each include a client identifier indicating a client from which the client originates, further comprising:

a queue selector that selects the queue from a plurality of queues based on the client identifier.

17. The system of claim 16, wherein the plurality of web scraping requests each further include an identifier of the target website, the queue selector is configured to select the queue from the plurality of queues based on the identifier of the target website.

18. The system of claim 17, further comprising the queue selector configured to select the queue from the plurality of queues based on whether the job description specifies that web scraping be made from a headless browser.

19. The system of claim 13, further comprising:

a throttle configuration database that stores a regular time window associated with the queue, wherein the job remover repeatedly removes a job description from the queue at the time window.

20. A non-transitory computer-readable device having instructions stored thereon that,

when executed by at least one computing device, cause the at least one computing device to perform operations comprising:

(a) during a first time period, receiving a plurality of web scraping requests from a client computing device, each of the plurality of web scraping requests specifying a target website to capture content from;

(b) placing job descriptions for the plurality of web scraping requests on a queue, the job descriptions each describing at least one web scraping job specified by a request from the plurality of requests;

(c) incrementally over a second time period longer than the first time period, removing the job descriptions from the queue;

when each respective job description is removed from the queue:

(d) generating a web request for the target website as specified in the job description; and

(e) transmitting the web request such that the web request reaches the target website via a proxy selected from a group of proxies.