Web Scraping Interview Questions and Answers

1. What is web scraping?

Web scraping is the technique to extract and read the data from the internet. The collected data can be saved and reused for data analytics.

2. Why web scraping?

There are many specific reasons why businesses may want to scrape a website; one of the vital reasons being the unavailability of APIs. Some of the other major reasons which may lead a company into scraping website are: Expand Market Share Due to the lack of availability of APIs the possibility of collaborating with business partners is limited. By exposing the data available in their website as APIs enterprises can open up new channels, possibilities to expand the market share and increase sales. Enter New Markets with Early go-to Market Strategy API being the long time strategy, Web Scraping solution can potentially enable organizations to build an early go-to market strategy. Access to Renewed and Structured Data Scraping the website of the organization through a Web Scraping solution gives an organization the chance to access renewed, structured and up to date data through the scraped APIs.

3. Explain Web Scraping Procedure.

There are multiple steps involved in web scraping:

- Reading data (source code of the web page URL) from the website
- Parsing this data based on the **HTML tags**
- Storing or displaying desired scraped information

Scraped data is very useful in data analytics.

4. What are the preferred programming languages for web scrapping?

Python is the most preferred programming language for web scrapping. It has many libraries to read and extract data from the internet, to parse and manipulate the data.
The data on the internet we access through the browser is in the HTML and CSS format. For extracting data from web pages, a **basic understanding of HTML tags** and CSS is required.

For storing data, JSON, XML, YAML formatting languages can be used.

4. Give an example of web scraping you worked on.

Extracting Stock Data Using Python library request, BeautifulSoup and panda html_read().

Note: You pick any examples and explain how you do it. You don't need to write complete code in an interview, but you have to explain the complete procedure and steps you followed. The interviewer can ask you many questions to test your knowledge.

5. What are the Python libraries you have used for web scrapping?

There are many Python libraries are available for web scrapping like…

- Beautiful Soap and Scrappy are the two most useful Python modules for scrapping web information.
- The request module is to read the data from internet web pages.
- JSON library is used to dump, to read and to write the JSON formatting objects.

- Beautiful Soup
- Requests
- Scrapy
- Selenium
- Urllib3

6. What is the purpose of the request module in Python?

The request module is used to read the data from the internet web pages. You have to pass the URL from where you want to read the data along with the HTTP request method, header information like encoding method, response data format, and session cookies…

In the HTTP response, you get data from the website. Data can be in any format like string, JSON, XML and YAML; based on data format mentioned in the request and server response.

7. What are the different HTTP response status codes?

When you send the HTTP request to read the data from the internet, you get the response along with the different response status.

Every status code has its meaning.

Sr. NO.	HTTP Method	Use
1	GET	It is used to access resources and to know the state of particular resource.
2	HEAD	This method requests the header which is required for client-server communication.
3	POST	It performs operations on a resource like creating and updating resource properties.
4	PUT	It is similar to the POST method. The only difference is, PUT follows the **idempotent** rule and POST does not.
5	DELETE	This method removes the resource.
6	OPTIONS	It describes the communication options for the target resource.

8. How to deal if your IP address is blocked by the website?

If you are accessing any website more than a certain threshold, your IP address can be blocked by the website. Proxy IPs/servers can be used to access the web pages if your IP address is blocked.

Usually, data analytics companies web scraps millions of web pages. Many times their IP addresses get blocked. To overcome this, they use a VPN (Virtual Private Network). There are many VPN service providers.

If you are not aware of VPN, here is how it works in laymen's terms.

How does VPN work?

You send a request to the VPN server. It reads the data from the website. VPN sends back the response to your IP address.

You can see, VPN actually hides your IP address from the website server and they will never come to know about your IP address. VPN has a pool of IP addresses. Even if the VPN IP address gets blocked, they can use another IP address from the pool.

9. What I can do with data scraping?

Web scraping is used for contact scraping, and as a component of applications used for web indexing, web mining and data mining, online price change monitoring and price comparison, product review scraping (to watch the competition), gathering real estate listings, weather data monitoring, website change detection, research, tracking online presence and reputation, web mashup and, web data integration.
Using data scraping you can build sitemaps that will navigate the site and extract the data. Using different type selectors, you will navigate the site and extract multiple types of data - text, tables, images, links and more.