1)What is Web Scraping? Why is it Used? Give three areas where Web Scraping is used to get data.

Ans- Web scraping is the process of collecting structured web data in an automated manner. It's also widely known as web data extraction or web data scraping. Some of the main use cases of web scraping include price monitoring, price intelligence, news monitoring, lead generation, and market research among many others.

Web scraping is the process of using bots to extract content and data from a website. Unlike screen scraping, which only copies pixels displayed onscreen, web scraping extracts underlying HTML code and, with it, data stored in a database. The scraper can then replicate entire website content elsewhere.

Web scraping typically extracts large amounts of data from websites for a variety of uses such as price monitoring, enriching machine learning models, financial data aggregation, monitoring consumer sentiment, news tracking, etc. Browsers show data from a website.

2) What are the different methods used for Web Scraping?

Ans-

Human copy-and-paste. The simplest form of web scraping is manually copying and pasting data from a web page into a text file or spreadsheet. ...

Text pattern matching. ...

HTTP programming. ...

HTML parsing. ...

DOM parsing. ...

Vertical aggregation. ...

Semantic annotation recognizing. ...

Computer vision web-page analysis.

3) What is Beautiful Soup? Why is it used?

Ans- Beautiful Soup is a Python package for parsing HTML and XML documents (including having malformed markup, i.e. non-closed tags, so named after tag soup). It creates a parse tree for parsed pages that can be used to extract data from HTML, which is useful for web scraping.

Beautiful Soup provides simple methods for navigating, searching, and modifying a parse tree in HTML, XML files. It transforms a complex HTML document into a tree of Python objects. It also automatically converts the document to Unicode, so you don't have to think about encodings.

4) Why is flask used in this Web Scraping project?

Ans- Flask is a lightweight framework to build websites. We'll use this to parse our collected data and display it as HTML in a new HTML file. The requests module allows us to send http requests to the website we want to scrape. The first line imports the Flask class and the render\_template method from the flask library.

5) Write the names of AWS services used in this project. Also, explain the use of each service.

Ans-

1. Amazon EC2

You don’t have to invest in costly physical services. Instead, you can create virtual machines with Amazon EC2 while managing other server features such as ports, security, and storage. Spend less time maintaining your servers and more time on your strategic projects. Invariably, Amazon EC2 is one of the most popular and fastest-growing of the many AWS services.

2. Amazon RDS

The Amazon Relational Database Service (RDS) was designed to make your infrastructure more user friendly. By using this AWS service, you can create dedicated instances of databases within minutes. Not to mention, these instances can support multiple database engines including SQL Server, SQL, PostgreSQL, and more. Take your time back and stop spending hours maintaining your database servers. Let Amazon RDS do the work for you.

3. Amazon Simple Storage Service (S3)

We are living in the age of big data. Some call it the incessant data deluge. As a result, we need more storage than ever before. Amazon Simple Storage Service (S3) has come to the rescue. It makes sense why this would be included in our list of the top 10 most used AWS services. It offers a highly secure and redundant file storage service. It also stores data in three data centers within a specific region. And, there’s more. Amazon S3 also offers integrations to help prevent breaches by way of PCI-DSS, HIPAA/HITECH, and FedRAMP. You get data flexibility without almost zero latency.

4. Amazon CloudFront

This service helps to improve website speed and access to cloud-based data. CloudFront works as a Global Content Delivery Service (CDN) to deliver content efficiently to end users. You’ll notice a significant increase in web page loading speed with this service. It even pulls website static files from data centers throughout the world.

5. Amazon VPC

If you are ready to isolate your entire IT infrastructure from exposure, then the only way to do it is with Amazon VPS. This service creates a private virtual network that cannot be accessed by anyone or anything except the people and systems you authorize.

For instance, you can quickly create a VPC via the AWS Management Console. Immediately, route tables, security groups, IP ranges, and Subnets are created. In addition, you get advanced security features such as network access control lists and security groups. If you need more security, you can create dedicated instances that are isolated at the hardware level from your other AWS accounts.

6. Amazon SNS

This is an event-driven computing hub that alerts subscriber services to perform tasks automatically in response to specified triggers.

Send notifications to any users on any platform. Further, Amazon ANS offers integration with PHP, Node, Python, and more. So then, you won’t miss a crucial alert at any time.

7. AWS Beanstalk

Developers have to deal with enough headaches as it is. AWS Beanstalk was created to help developers manage website infrastructure. It’s difficult for developers to switch from development to maintenance at the drop of a hat. Yet, AWS Beanstalk offers autoscaling to ensure automatic updates of new software. And, this service runs automatically.

AWS Beanstalk really is a timesaver. It automates the setup, configuration, and provisioning of other AWS services such as EC2, RDS, and S3. Not to mention, the automated setup also helps to mitigate human error.

8. AWS Lambda

When your server is inundated with an influx of requests, are you overwhelmed without knowing precisely how to respond? It’s quite possible your current server infrastructure cannot support the demands of your current speed of development. In this case, AWS Lambda is designed to support any load of development. You handle the coding, and AWS Lambda will offer the right amounts of support and required resources while scaling to ensure your systems are no longer stretched past capacity.

9. AWS Autoscaling

If your organization is planning for any growth, then you need an IT infrastructure that can offer the necessary support. AWS Autoscaling can manage fleets of servers and the incoming traffic. Multiple instances are created when needed. It’s almost as if you can never grow large enough.

It also offers predictive scaling and provisions the perfect amount of EC2 instances ahead of future traffic such as seasonal spikes. Moreover, using machine learning algorithms, it can detect daily and weekly patterns and adjust accordingly.

10. AWS IAM

The last service listed on our top 10 most used AWS services is AWS Identity and Access Management (IAM). Unquestionably, security has a lot to do with access and what gets accessed. This service offers an effective fortification of sensitive data and AWS resources. It can also be used in conjunction with your organization’s 2FA and MFA. It’s just an additional layer of security that never hurts.

.